

Quasi-Bayesian analysis of nonparametric instrumental variables models^{*}

Kengo Kato

*Department of Mathematics
Graduate School of Science
Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima
Hiroshima 739-8526, Japan.
e-mail: kkato@hiroshima-u.ac.jp*

Abstract: This paper aims at developing a quasi-Bayesian analysis of the nonparametric instrumental variables model, with a focus on the asymptotic properties of quasi-posterior distributions. In this paper, instead of assuming a distributional assumption on the data generating process, we consider a quasi-likelihood induced from the conditional moment restriction, and put priors on the function-valued parameter. We call the resulting posterior quasi-posterior, which corresponds to “Gibbs posterior” in the literature. Here we use sieve priors, which are prior distributions that concentrate on finite dimensional sieve spaces. The dimension of the sieve space should increase as the sample size. We derive rates of contraction and a non-parametric Bernstein-von Mises type result for the quasi-posterior distribution, and rates of convergence for the quasi-Bayes estimator defined by the posterior expectation. We show that, with priors suitably chosen, the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, respectively). These results greatly sharpen the previous related work.

AMS 2000 subject classifications: 62G08, 62G20.

Keywords and phrases: asymptotic normality, inverse problem, nonparametric instrumental variables model, quasi-Bayes, rates of contraction.

1. Introduction

1.1. Overview

Let (Y, X, W) be a triplet of scalar random variables, where Y is a dependent variable, X is an endogenous variable and W is an instrumental variable. Without losing much generality, we assume that the support of (X, W) is contained in $[0, 1]^2$. The support of Y may be unbounded. We consider the nonparametric instrumental variables (NPIV) model of the form

$$\mathbb{E}[Y \mid W] = \mathbb{E}[g_0(X) \mid W], \quad (1)$$

^{*}Supported by the Grant-in-Aid for Young Scientists (B) (22730179) from the JSPS.

where $g_0 : [0, 1] \rightarrow \mathbb{R}$ is an unknown structural function of interest. If we define $U = Y - g_0(X)$, (1) reduces to the conventional form

$$Y = g_0(X) + U, \quad \mathbb{E}[U \mid W] = 0.$$

Here X is potentially correlated with U and hence $\mathbb{E}[U \mid X] \neq 0$.

Suppose that (X, W) has square-integrable joint density $f_{X,W}(x, w)$ on $[0, 1]^2$ and denote by $f_W(w)$ the density of W . Define the linear operator $K : L_2[0, 1] \rightarrow L_2[0, 1]$ by

$$(Kg)(w) = \mathbb{E}[g(X) \mid W = w]f_W(w) = \int g(x)f_{X,W}(x, w)dx.$$

Let $h(w) = \mathbb{E}[Y \mid W = w]f_W(w)$. Then, the conditional moment restriction (1) is equivalent to the operator equation

$$Kg_0 = h. \quad (2)$$

Assume that K is injective to guarantee identification of g_0 . The function h is relatively standard to estimate. However, even though K is injective, its inverse K^{-1} is not L_2 -continuous since K is Hilbert-Schmidt and hence the l -th largest singular value, denoted by κ_l , is approaching zero as $l \rightarrow \infty$. In this sense, the problem of recovering g_0 from h is *ill-posed*.

A model of the form (1) is of principal importance in econometrics (see [Hall and Horowitz, 2005](#); [Horowitz, 2011](#)). From a statistical perspective, the problem of recovering the structural function g_0 is challenging since it is an ill-posed inverse problem with an additional difficulty of *unknown* K (furthermore, it is not plausible to think of that K is known up to a random error independent of the data, which is a notable difference from the case considered in [Hoffman and Reiss, 2008](#)). Statistical inverse problems, including the current problem, have attracted considerable interest in statistics, econometrics and mathematical analysis. We refer the reader to [Kless \(1999\)](#) for a textbook treatment of linear inverse problems, and [Cavalier \(2008\)](#) for a recent review on statistical inverse problems.

Approaches to estimating the structural function g_0 are roughly classified into two types: the method involving the Tikhonov regularization ([Hall and Horowitz, 2005](#); [Darolles et al., 2011](#)) and the sieve-based method ([Newey and Powell, 2003](#); [Ai and Chen, 2003](#); [Blundell et al., 2007](#); [Horowitz, 2012](#)).¹ The minimax optimal rates of convergence in estimating the structural function g_0 are established in [Hall and Horowitz \(2005\)](#) and [Chen and Reiss \(2011\)](#). Similarly to other statistical inverse problems, these rates are characterized by the smoothness of g_0 and the “ill-posedness” of the problem. The optimal rates are achieved by the estimators proposed by [Hall and Horowitz \(2005\)](#) and [Blundell et al. \(2007\)](#) under their respective assumptions.

All the above mentioned studies are from a purely frequentist perspective. Little is known about the theoretical properties of Bayes or quasi-Bayes analysis of

¹The sieve-based method is approximately the Galerkin projection method in mathematical analysis.

the NPIV model. Exceptions are [Florens and Simoni \(2012a\)](#) and [Liao and Jiang \(2011\)](#) with which we will compare our work in the subsection below.

This paper aims at developing a quasi-Bayesian analysis of the NPIV model, with a focus on the asymptotic properties of quasi-posterior distributions. The approach taken is quasi-Bayes in the sense that any specific distribution of (Y, X, W) is not assumed and the analysis is based upon a quasi-likelihood induced from the conditional moment restriction. The quasi-likelihood is constructed by first estimating the conditional moment function $m(\cdot, g) = \mathbb{E}[Y - g(X) \mid W = \cdot]$ in a nonparametric way, and taking $\exp\{-(1/2) \sum_{i=1}^n \hat{m}^2(W_i, g)\}$ as if it were a likelihood of g . For this quasi-likelihood, we put a prior on the function-valued parameter g . By doing so, formally, the posterior distribution for g may be defined, which we call “quasi-posterior distribution”. This posterior corresponds to what [Jiang and Tanner \(2008\)](#) called “Gibbs posterior”, and has a substantial interpretation (see Proposition 1 ahead).

This framework of the quasi-posterior (Gibbs posterior) allows us a flexibility since a stringent distributional assumption, such as normality, on the data generating process is not required. Such a framework widens a Bayesian approach to broad fields of statistical problems, as [Jiang and Tanner \(2008, p.2211\)](#) remarked: “This framework of the Gibbs posterior has been overlooked by most statisticians for a long time [...] a foundation for understanding the statistical behavior of the Gibbs posterior, which we believe will open a productive new line of research.”

In this paper, we shall use sieve priors, which are prior distributions that concentrate on finite dimensional sieve spaces. The dimension of the sieve space, which plays a role of regularization parameter, should go to infinity as the sample size. Potentially, there are several choices in sieve spaces. Here we choose to use wavelet bases to form sieve spaces. Wavelet bases are useful to treat smoothness function classes such as Hölder-Zygmund and Sobolev spaces in a unified and convenient way. Likewise, we shall use wavelet series estimation of the conditional moment function $m(\cdot, g)$.²

Under this setup, we study the asymptotic properties of the quasi-posterior distribution. The results obtained are summarized as follows. First, we derive rates of contraction for the quasi-posterior distribution and establish conditions on priors under which the minimax optimal rate of contraction is attained. Here the contraction is stated in the standard L_2 -norm. Second, we show asymptotic normality of the quasi-posterior of the first k_n generalized Fourier coefficients, where $k_n \rightarrow \infty$ is the dimension of the sieve space. This may be viewed as a nonparametric Bernstein-von Mises type result (see [van der Vaart, 1998](#), Chapter 10 for the classical Bernstein-von Mises theorem for regular parametric models). Third, we derive rates of convergence of the quasi-Bayes estimator defined by the posterior expectation and show that under some conditions it attains the minimax optimal rate of convergence. Finally, we give some specific sieve priors for which the quasi-posterior distribution (the quasi-Bayes estimator) attains the

²This does not rule out the use of other bases such as the Fourier and Hermite polynomial bases. See Remark 5.

minimax optimal rate of contraction (convergence, respectively). These results greatly sharpen the previous work of e.g. [Liao and Jiang \(2011\)](#), as we will review below.

1.2. Literature review and contributions

Closely related are [Florens and Simoni \(2012a\)](#) and [Liao and Jiang \(2011\)](#). The former paper worked on the reduced form equation $Y = \mathbb{E}[g_0(X) | W] + V$ with $V = U + g_0(X) - \mathbb{E}[g_0(X) | W]$ and assumed V to be normally distributed. They considered a Gaussian prior on g , and because of that the posterior distribution is also Gaussian. They proposed to “regularize” the posterior and established frequentist rates for the “regularized” posterior mean. Obviously, the present paper largely differs from [Florens and Simoni \(2012a\)](#) in that (i) we do not assume normality of the “error”; (ii) roughly speaking, Florens and Simoni’s method is tied with the Tikhonov regularization method, while ours is tied with the sieve-based method. [Liao and Jiang \(2011\)](#) developed an important unified framework in estimating conditional moment restriction models based on a quasi-Bayesian approach, and their scope is more general than ours. They analyzed NPIV models in detail in their Section 4. Their posterior construction is similar to ours such as the use of sieve priors, but differs from ours in detail. For example, [Liao and Jiang \(2011\)](#) transformed the conditional moment restriction into unconditional moment restrictions with increasing number of restrictions. On the other hand, we directly work on the conditional moment restriction.

Importantly and substantially, none of these papers did not establish sharp contraction rates for their (quasi-)posterior distributions, nor asymptotic normality results. It is unclear whether Florens and Simoni’s rates are optimal, since their assumptions are substantially different from the past literature such as [Hall and Horowitz \(2005\)](#) and [Chen and Reiss \(2011\)](#). Liao and Jiang only established posterior consistency. Here, we focus on a simple but important model, and establish the sharper asymptotic results for the quasi-posterior distribution. Notably, a wide class of (finite dimensional) sieve priors is shown to lead to the optimal contraction rate. Moreover, in [Liao and Jiang \(2011\)](#), a point estimator of the structural function is not formally analyzed. Hence the primal contribution of this paper is to considerably deepen the understanding of the asymptotic properties of the quasi-Bayesian procedure for the NPIV model.

The present paper deals with a quasi-Bayesian analysis of an infinite dimensional model. The literature on theoretical studies of Bayesian analysis of infinite dimensional models is large. [Ghosh and Ramamoorthi \(2003\)](#) is a good reference on this topic. We refer the reader to [Ghosal et al. \(2000\)](#); [Shen and Wasserman \(2001\)](#); [Kleijn and van der Vaart \(2006\)](#); [Ghosal and van der Vaart \(2007\)](#) for general contraction rates results for posterior distributions in infinite dimensional models. Note that these results do not directly apply to our case since the “likelihood” here is nonparametrically estimated. The paper also contributes to the literature on Bayesian analysis of linear inverse problems, which we believe is still an under-developed area of research. For nonparametric Bayesian

analysis of inverse problems other than NPIV models, we refer to [Cox \(1993\)](#); [Florens and Simoni \(2012b\)](#); [Knapik et al. \(2011\)](#).

Our asymptotic normality result builds upon the previous work on asymptotic normality of (quasi-)posterior distributions for models with increasing number of parameters ([Ghosal, 1999, 2000](#); [Belloni and Chernozhukov, 2009a,b](#); [Boucheron and Gassiat, 2009](#); [Clarke and Ghosal, 2010](#); [Bontemps, 2011](#)). Related is [Bontemps \(2011\)](#), in which the author established Bernstein-von Mises theorems for Gaussian regression models with increasing number of regressors and improved upon the earlier work of [Ghosal \(1999\)](#) in several aspects. [Bontemps \(2011\)](#) covered nonparametric models by taking into account modeling bias in the analysis. However, none of these papers did not cover the NPIV model, nor more generally linear inverse problems.

1.3. Organization and notation

The remainder of the paper is organized as follows. Section 2 gives an informal discussion of the quasi-Bayesian analysis of the NPIV model. Section 3 summarizes some basic facts on wavelet theory and introduces the posterior construction used in the analysis. Section 4 contains the main results of the paper. Section 5 analyzes some specific sieve priors. Section 6 contains the proofs of the main results. Section 7 concludes with some further discussions. Appendix contains some technical results omitted in the main body.

Notation: For any given (random or non-random, scalar or vector) sequence $\{z_i\}_{i=1}^n$, we use the notation

$$\mathbb{E}_n[z_i] = n^{-1} \sum_{i=1}^n z_i,$$

which should be distinguished from the population expectation $\mathbb{E}[\cdot]$. For any vector z , let $z^{\otimes 2} = zz^T$ where z^T is the transpose of z . For any two sequences of positive constants r_n and s_n , we write $r_n \lesssim s_n$ if the ratio r_n/s_n is bounded, and $r_n \sim s_n$ if $r_n \lesssim s_n$ and $s_n \lesssim r_n$. Let $L_2[0, 1]$ denote the usual L_2 space with respect to the Lebesgue measure for functions defined on $[0, 1]$. Let $\|\cdot\|$ denote the L_2 -norm, i.e., $\|f\|^2 = \int_0^1 f^2(x)dx$. The inner product in $L_2[0, 1]$ is denoted by $\langle \cdot, \cdot \rangle$, i.e., $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$. Let $C[0, 1]$ denote the metric space of all continuous functions on $[0, 1]$, equipped with the uniform metric. For any function $f : [0, 1] \rightarrow \mathbb{R}$, let $\|f\|_\infty = \sup_{x \in [0, 1]} |f(x)|$. The Euclidean norm is denoted by $\|\cdot\|_{\ell^2}$. For any matrix A , let $s_{\min}(A)$ and $s_{\max}(A)$ denote the minimum and maximum singular values of A , respectively. Let $\|A\|_{\text{op}}$ denote the operator norm of matrix A (i.e., $\|A\|_{\text{op}} = s_{\max}(A)$). Denote by $dN(\mu, \Sigma)(x)$ the density of the multivariate normal distribution with mean vector μ and covariance matrix Σ .

2. Quasi-Bayesian analysis: informal discussion

In this section, we outline a quasi-Bayesian analysis of the NPIV model (1). The discussion here is informal. The formal discussion is given in Section 4.

Let \mathcal{G} be a parameter space (say, some smoothness class of functions, such as a Hölder-Zygmund or Sobolev space), for which we assume $g_0 \in \mathcal{G}$. We assume that \mathcal{G} is at least contained in $C[0, 1]$: $\mathcal{G} \subset C[0, 1]$. Define the conditional moment function as $m(W, g) = \mathbb{E}[Y - g(X) \mid W]$, $g \in \mathcal{G}$. Then g_0 satisfies the conditional moment restriction

$$m(W, g_0) = 0, a.s. \quad (3)$$

Equivalently, we have $\mathbb{E}[m^2(W, g_0)] = 0$.

In this paper, any specific distribution of (Y, X, W) is not assumed. So a Bayesian analysis in the standard sense is not applicable here since a proper likelihood for g (g is a generic version of g_0) is not available. Instead, we use a quasi-likelihood induced from the conditional moment restriction (3).

Let $(Y_1, X_1, W_1), \dots, (Y_n, X_n, W_n)$ be i.i.d. observations of (Y, X, W) . Let $W^n = \{W_1, \dots, W_n\}$ and $\mathcal{D}_n = \{(Y_1, X_1, W_1), \dots, (Y_n, X_n, W_n)\}$. By (3), a plausible candidate of the quasi-likelihood would be

$$p_g(W^n) = \exp\{-(n/2)\mathbb{E}_n[m^2(W_i, g)]\},$$

since $p_g(W^n)$ is maximized at the true structural function g_0 . Here, recall that $\mathbb{E}_n[z_i] = n^{-1} \sum_{i=1}^n z_i$ for any sequence $\{z_i\}_{i=1}^n$. However, this $p_g(W^n)$ is infeasible since $m(\cdot, g)$ is unknown. Instead of using $p_g(W^n)$, we replace $m(\cdot, g)$ by a suitable estimate $\hat{m}(\cdot, g)$ and use the quasi-likelihood of the form

$$p_g(\mathcal{D}_n) = \exp\{-(n/2)\mathbb{E}_n[\hat{m}^2(W_i, g)]\}.$$

Below we use a wavelet series estimator of $m(\cdot, g)$.

The quasi-Bayesian analysis considered here uses this quasi-likelihood as if it were a proper likelihood and puts priors on $g \in \mathcal{G}$. In this paper, as in [Liao and Jiang \(2011\)](#), we shall use sieve priors. The basic idea is to construct a sequence of finite dimensional sieve spaces (say, \mathcal{G}_n) that well approximates the parameter space \mathcal{G} (i.e., each function in \mathcal{G} is well approximated by some function in \mathcal{G}_n as n becomes large), and put priors concentrating on these sieve spaces. Each sieve space is a subset of a linear space spanned by some basis functions. Hence the problem reduces to putting priors on the coefficients on those basis functions. Such priors are typically called “(finite dimensional) sieve priors” (or “series priors”) and have been widely used in the nonparametric Bayesian and quasi-Bayesian analysis (see e.g. [Ghosal et al., 2000](#); [Scricciolo, 2006](#); [Ghosal and van der Vaart, 2007](#)).

Let Π_n be a so-constructed prior on $g \in \mathcal{G}$. Formally, the posterior distribution of g given \mathcal{D}_n may be defined by

$$\Pi_n(dg \mid \mathcal{D}_n) = \frac{p_g(\mathcal{D}_n)\Pi_n(dg)}{\int p_g(\mathcal{D}_n)\Pi_n(dg)}, \quad (4)$$

which we call “quasi-posterior distribution”. The quasi-posterior distribution is not a proper posterior distribution in the strict Bayesian sense since $p_g(\mathcal{D}_n)$ is not a proper likelihood. Nevertheless, $\Pi_n(dg \mid \mathcal{D}_n)$ is a proper distribution, i.e., $\int \Pi_n(dg \mid \mathcal{D}_n) = 1$. Similarly to proper posterior distributions, contraction of the quasi-posterior distribution around g_0 intuitively means that it contains more and more accurate information about the true structural function g_0 as the sample size increases. Hence, as in proper posterior distributions, it is of fundamental importance to study rates of contraction of quasi-posterior distributions. Here we say that the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ contracts around g_0 at rate $\varepsilon_n \rightarrow 0$ if $\Pi_n(g : \|g - g_0\| > \varepsilon_n \mid \mathcal{D}_n) \xrightarrow{P} 0$.

This quasi-posterior corresponds to what [Zhang \(2006b\)](#) called “Gibbs algorithm” and what [Jiang and Tanner \(2008\)](#) called “Gibbs posterior”. Here an interesting interpretation of the quasi-posterior is obtained.

Proposition 1. *Let $\eta > 0$ be a fixed constant. Let Π be a prior distribution for g defined on, say, the Borel σ -field of $C[0, 1]$. Suppose that the data \mathcal{D}_n are fixed and the maps $g \mapsto \hat{m}_i(W_i, g)$ are measurable with respect to the Borel σ -field of $C[0, 1]$. Then, the distribution*

$$\hat{\Pi}_\eta(dg) = \frac{\exp(-\eta \sum_{i=1}^n \hat{m}_i^2(W_i, g)) \Pi(dg)}{\int \exp(-\eta \sum_{i=1}^n \hat{m}_i^2(W_i, g)) \Pi(dg)},$$

minimizes the empirical information complexity defined by

$$\check{\Pi} \mapsto \int \sum_{i=1}^n \hat{m}_i^2(W_i, g) \check{\Pi}(dg) + \eta^{-1} D_{KL}(\check{\Pi} \parallel \Pi) \quad (5)$$

over all distributions $\check{\Pi}$ absolutely continuous with respect to Π . Here

$$D_{KL}(\check{\Pi} \parallel \Pi) = \int \check{\pi} \log \check{\pi} \Pi(dg), \text{ with } d\check{\Pi}/d\Pi = \check{\pi},$$

is the Kullback-Leibler divergence from $\check{\Pi}$ to Π .

Proof. Immediate from [Zhang \(2006a, Proposition 5.1\)](#). □

The proposition shows that, given the data \mathcal{D}_n and a prior $\Pi = \Pi_n$ on g , the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ defined in (4) is obtained as a minimizer of the empirical information complexity defined by (5) with $\eta = 1/2$. This gives a rationale to use $\Pi_n(dg \mid \mathcal{D}_n)$ as a quasi-posterior since, among all possible “quasi-posteriors”, this $\Pi_n(dg \mid \mathcal{D}_n)$ optimally balances the average of the natural loss function $g \mapsto \sum_{i=1}^n \hat{m}_i^2(W_i, g)$ and its complexity (or deviation) relative to the initial prior distribution measured by the Kullback-Leibler divergence. The scaling constant (“temperature”) η is taken to be $1/2$ here. However, changing this value does not substantially affect the asymptotic analysis.

The quasi-posterior distribution provides point estimators of g_0 . A most natural estimator would be the estimator defined by the posterior expectation (the

expectation of the quasi-posterior distribution), i.e.,

$$\hat{g}_{QB} = \begin{cases} \int g \Pi_n(dg \mid \mathcal{D}_n), & \text{if the right integral exists,} \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where the integral $\int g \Pi_n(dg \mid \mathcal{D}_n)$ is understood as pointwise.

In Section 4, we will study the asymptotic properties of the quasi-posterior distribution and the quasi-Bayes estimator. In doing so, we have to specify certain regularity properties, such as the smoothness of g_0 and the degree of ill-posedness of the problem. How to characterize the “smoothness” of g_0 is important here since it is related to how to put priors. For that purpose, we find wavelet theory useful, and use sieve spaces constructed by using wavelet bases.

3. Wavelets, function spaces and posterior construction

3.1. Wavelet bases for $L_2[0, 1]$

We review wavelet theory on the compact interval $[0, 1]$. We refer the reader to [Härdle et al. \(1998\)](#), [Mallat \(2009\)](#) and [Johnstone \(2011, Chapter 7 and Appendix B\)](#) as useful general references on wavelet theory in the statistical (and signal processing) context.

Let (φ, ψ) be a Daubechies pair of the scaling function and wavelet of a multiresolution analysis of the space $L_2(\mathbb{R})$ of order N , with ψ having N vanishing moments and support contained in $[-N+1, N]$, and φ having support contained in $[0, 2N-1]$ (see [Härdle et al., 1998](#), Remark 7.1). We translate φ so that its support is contained in $[-N+1, N]$. Define

$$\varphi_{jk}(x) = 2^{j/2} \varphi(2^j x - k), \quad \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k).$$

Then, for any fixed $J_0 \geq 0$, it is known that $\{\varphi_{J_0 k}, \psi_{jk}, j \geq J_0, k \in \mathbb{Z}\}$ forms an orthonormal basis for $L_2(\mathbb{R})$. However, we need an orthonormal basis for $L_2[0, 1]$. From the Daubechies pair (φ, ψ) , we wish to construct an orthonormal basis for $L_2[0, 1]$. The construction here is based on [Cohen et al. \(1993, Section 4\)](#). See also Chapter 7.5 of [Mallat \(2009\)](#) for wavelet bases on $[0, 1]$.

Take a fixed resolution level j such that $2^j \geq 2N$. For $k = N, \dots, 2^j - N - 1$, φ_{jk} are supported in $[0, 1]$ and left unchanged: $\varphi_{jk}^{\text{int}}(x) = \varphi_{jk}(x)$ for $x \in [0, 1]$. At boundaries, $k = 0, \dots, N - 1$, construct *some* functions φ_k^L with support $[0, N + k]$ and φ_k^R with support $[-N - k, 0]$, and define

$$\varphi_{jk}^{\text{int}}(x) = 2^{j/2} \varphi_k^L(2^j x), \quad \varphi_{j, 2^j - k - 1}^{\text{int}}(x) = 2^{j/2} \varphi_k^R(2^j(x - 1)), \quad x \in [0, 1].$$

Note that both φ_k^L and φ_k^R have the same smoothness as φ . Define the multiresolution spaces $V_j = \text{span}\{\varphi_{jk}^{\text{int}}, k = 0, \dots, 2^j - 1\}$, which satisfy the following properties (i) $\dim(V_j) = 2^j$; (ii) $V_j \subset V_{j+1}$; (iii) each V_j contains all polynomials of order at most $N - 1$.

Turning to the wavelet spaces, define W_j by the orthogonal complement of V_j in V_{j+1} . Starting from the Daubechies wavelet ψ , construct ψ_{jk}^{int} similarly to $\varphi_{jk}^{\text{int}}$. Then, we have $W_j = \text{span}\{\psi_{jk}^{\text{int}}, k = 0, \dots, 2^j - 1\}$, and for any $J_0 \geq 1$ with $2^{J_0} \geq 2N$ and $J > J_0$,

$$V_J = V_{J_0} \bigoplus_{j \geq J_0}^{J-1} W_j, \quad L_2[0, 1] = V_{J_0} \bigoplus_{j \geq J_0} W_j.$$

Therefore, $\{\varphi_{J_0 k}^{\text{int}}\}_{k=0}^{2^{J_0}-1} \cup \{\psi_{jk}^{\text{int}}, j \geq J_0, k = 0, \dots, 2^j - 1\}$ forms an orthonormal basis for $L_2[0, 1]$ (see Section 4 of [Cohen et al., 1993](#), for formal proofs of these results)

To make the notation simpler, define functions ϕ_1, ϕ_2, \dots by

$$\begin{aligned} \phi_1 &= \varphi_{J_0, 0}^{\text{int}}, \phi_2 = \varphi_{J_0, 1}^{\text{int}} \dots, \phi_{2^{J_0}} = \varphi_{J_0, 2^{J_0}-1}^{\text{int}}, \\ \phi_{2^{J_0}+1} &= \psi_{J_0, 0}^{\text{int}}, \phi_{2^{J_0}+2} = \psi_{J_0, 1}^{\text{int}}, \dots, \phi_{2^{J_0}+1} = \psi_{J_0, 2^{J_0}-1}^{\text{int}}, \\ \phi_{2^{J_0}+1+1} &= \psi_{J_0+1, 0}^{\text{int}}, \phi_{2^{J_0}+1+2} = \psi_{J_0+1, 1}^{\text{int}}, \dots, \phi_{2^{J_0}+2} = \psi_{J_0+1, 2^{J_0}-1}^{\text{int}}, \\ &\vdots \\ \phi_{2^j+1} &= \psi_{j, 0}^{\text{int}}, \phi_{2^j+2} = \psi_{j, 1}^{\text{int}}, \dots, \phi_{2^j+1} = \psi_{j, 2^j-1}^{\text{int}}, \\ &\vdots \end{aligned}$$

so that we have

$$\{\varphi_{J_0 k}^{\text{int}}\}_{k=0}^{2^{J_0}-1} \cup \{\psi_{jk}^{\text{int}}, j \geq J_0, k = 0, \dots, 2^j - 1\} = \{\phi_l, l \geq 1\}.$$

Note that

$$V_j = \text{span}\{\phi_1, \dots, \phi_{2^j}\}, \quad j \geq J_0.$$

Definition 1. Call the so-constructed basis $\{\phi_l, l \geq 1\}$ the CDV (Cohen-Daubechies-Vial) wavelet basis for $L_2[0, 1]$ generated from the Daubechies pair (φ, ψ) . If (φ, ψ) is S -regular, i.e., if (φ, ψ) are S -times continuously differentiable, then call the so-generated CDV wavelet basis $\{\phi_l, l \geq 1\}$ S -regular.

Remark 1. For any given positive integer S , there is an S -regular Daubechies pair (φ, ψ) by taking the order N sufficiently large (see [Härdle et al., 1998](#), Remark 7.1).

Finally, denote by P_j the projection operator from $L_2[0, 1]$ onto the j -th multiresolution space V_j , i.e., for any $g = \sum_{l=1}^{\infty} b_l \phi_l \in L_2[0, 1]$, $P_j g = \sum_{l=1}^{2^j} b_l \phi_l$.

In what follows, for any $J \in \mathbb{N}$, the notation of kind b^J means that it is a vector of dimension 2^J . For example, $b^J = (b_1, \dots, b_{2^J})^T$.

3.2. Function spaces

We recall the definition of Besov spaces.

Definition 2. Let $0 < s < S$, $s \in \mathbb{R}$, $S \in \mathbb{N}$ and $1 \leq p, q \leq \infty$. Let $\{\phi_l, l \geq 1\}$ be an S -regular CDV wavelet basis for $L_2[0, 1]$. Denote by $b_l(f) = \int f \phi_l$ the generalized Fourier coefficients of $f \in L_2[0, 1]$. Then the Besov space $B_{p,q}^s$ is defined by the set of functions $\{f \in L_2[0, 1] : \|f\|_{s,p,q} < \infty\}$, where

$$\|f\|_{s,p,q} := \left(\sum_{1 \leq k \leq 2^{J_0}} |b_k(f)|^p \right)^{1/p} + \left(\sum_{j \geq J_0} \left(2^{j(s+1/2-1/p)} \left(\sum_{1 \leq k \leq 2^j} |b_{2^j+k}(f)|^p \right)^{1/p} \right)^q \right)^{1/q},$$

with the obvious modification in case $p = \infty$ or $q = \infty$.

Remark 2. Besov spaces cover commonly used smooth function spaces. For example, $B_{\infty,\infty}^s$ is equal to the Hölder-Zygmund space, which coincides with the classical Hölder space for non-integer s . For integer s , they do not coincide but the Hölder-Zygmund space contains the classical Hölder space. Furthermore, $B_{2,2}^s$ is equal to the classical L_2 -Sobolev space when s is an integer.

Remark 3 (Approximation property). For either $g \in B_{\infty,\infty}^s$ or $B_{2,2}^s$, we have $\|g - P_J g\|^2 \leq C 2^{-2Js}$ for all $J \geq J_0$. Here the constant C depends only on s and the corresponding Besov norm of g .

As a parameter space, we assume $\mathcal{G} = \mathcal{G}^s = B_{\infty,\infty}^s$ (Hölder-Zygmund) or $B_{2,2}^s$ (Sobolev) for some $s > 1/2$. Note that in either case $\mathcal{G}^s \subset C[0, 1]$.

In what follows,

take and fix an S -regular CDV wavelet basis $\{\phi_l, l \geq 1\}$ with $S > s$.

We keep this convention throughout the analysis.

3.3. Posterior construction

To construct quasi-posterior distributions, we have to estimate $m(\cdot, g)$ and construct a sequence of sieve spaces for \mathcal{G}^s on which priors concentrate. For the former purpose, we use a wavelet series estimator of $m(\cdot, g)$. For the latter purpose, we construct a sequence of sieve spaces formed by the wavelet basis.

For $J \geq J_0$, define the 2^J -dimensional vector of functions $\phi^J(w)$ by

$$\phi^J(w) = (\phi_1(w), \dots, \phi_{2^J}(w))^T.$$

Let $J_n \geq J_0$ be a sequence of positive integers such that $J_n \rightarrow \infty$ and $2^{J_n} = o(n)$. Let

$$\hat{m}(w, g) = \phi^{J_n}(w)^T (\mathbb{E}_n[\phi^{J_n}(W_i)^{\otimes 2}])^{-1} \mathbb{E}_n[\phi^{J_n}(W_i)(Y_i - g(X_i))],$$

which is a wavelet series estimator of $m(\cdot, g)$ (replace the inverse matrix by the generalized inverse if the former does not exist; the probability of such an event converges to zero as $n \rightarrow \infty$ under the assumptions below). We use this wavelet series estimator throughout the analysis (see the remark at the end of the section).

For the same J_n , we shall take $V_{J_n} = \text{span}\{\phi_1, \dots, \phi_{2^{J_n}}\}$ as a sieve space for \mathcal{G}^s . We consider priors Π_n that concentrate on V_{J_n} , i.e., $\Pi_n(V_{J_n}) = 1$. Formally, we think of that priors on g are defined on the Borel σ -field of $C[0, 1]$ (hence the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ is understood to be defined on the Borel σ -field of $C[0, 1]$, which is possible since the map $g \mapsto p_g(\mathcal{D}_n)$ here is continuous on $C[0, 1]$). Since the map $b^{J_n} = (b_1, \dots, b_{2^{J_n}})^T \mapsto \sum_{l=1}^{2^{J_n}} b_l \phi_l, \mathbb{R}^{2^{J_n}} \rightarrow C[0, 1]$, is homeomorphic from $\mathbb{R}^{2^{J_n}}$ onto V_{J_n} , putting priors on $g \in V_{J_n}$ is equivalent to putting priors on $b^{J_n} \in \mathbb{R}^{2^{J_n}}$ (the latter are of course defined on the Borel σ -field of $\mathbb{R}^{2^{J_n}}$). Practically, priors on $g \in V_{J_n}$ are induced from priors on $b^{J_n} \in \mathbb{R}^{2^{J_n}}$. For the later purpose, it is useful to determine the correspondence between priors for these two parameterizations. Unless otherwise stated, we follow the convention of the notation such that:

$$\tilde{\Pi}_n: \text{a prior on } b^{J_n} \in \mathbb{R}^{2^{J_n}} \leftrightarrow \Pi_n: \text{the induced prior on } g \in V_{J_n}.$$

We shall call $\tilde{\Pi}_n$ a generating prior, and Π_n the induced prior.

Correspondingly, the quasi-posterior for b^{J_n} is defined. With a slight abuse of notation, for $g = \sum_{l=1}^{2^{J_n}} b_l \phi_l$, we write $\hat{m}(w, b^{J_n}) = \hat{m}(w, g)$, and take $p_{b^{J_n}}(\mathcal{D}_n) = \exp\{-(n/2)\mathbb{E}_n[\hat{m}^2(W_i, b^{J_n})]\}$ as a quasi-likelihood for b^{J_n} . Note that in this particular setting, the log quasi-likelihood is quadratic in b^{J_n} . Let $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$ denote the resulting quasi-posterior distribution for b^{J_n} , i.e.,

$$\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n) = \frac{p_{b^{J_n}}(\mathcal{D}_n) \tilde{\Pi}_n(db^{J_n})}{\int p_{b^{J_n}}(\mathcal{D}_n) \tilde{\Pi}_n(db^{J_n})}. \quad (7)$$

For the quasi-Bayes estimator \hat{g}_{QB} defined by (6), since for every $x \in [0, 1]$, the map $g \mapsto g(x)$ is continuous on $C[0, 1]$, and conditional on \mathcal{D}_n the quasi-posterior $\Pi_n(dg \mid \mathcal{D}_n)$ is a Borel probability measure on $C[0, 1]$, the integral $\int g(x) \Pi_n(dg \mid \mathcal{D}_n)$ exists as soon as $\int |g(x)| \Pi_n(dg \mid \mathcal{D}_n) < \infty$. Furthermore, \hat{g}_{QB} can be computed by using the relation

$$\int g(x) \Pi_n(dg \mid \mathcal{D}_n) = \phi^{J_n}(x)^T \left[\int b^{J_n} \tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n) \right],$$

as soon as one of the integral on the right side exists. Hence, practically, it is sufficient to compute the expectation of $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$.

Remark 4. The use of the same wavelet basis to estimate $m(\cdot, g)$ and to construct a sequence of sieve spaces for \mathcal{G}^s is not essential and can be relaxed. Suppose that we have another CDV wavelet basis $\{\tilde{\phi}_l\}$ for $L_2[0, 1]$ and use this basis to estimate $m(\cdot, g)$. Then, all the results below apply by simply replacing $\phi_l(W_i)$ by $\tilde{\phi}_l(W_i)$. To keep the notation simple, we use the same wavelet basis.

However, the use of the same resolution level J_n is essential (at least at the proof level) in establishing the asymptotic properties of the quasi-posterior distribution. It may be a technical artifact, but we do not extend the theory in this direction since there is no clear theoretical benefit to do so.

Remark 5. The use of CDV wavelet bases is not crucial and one may use other reasonable bases such as the Fourier and Hermite polynomial bases. The theory below can be extended to such bases with some modifications. However, CDV wavelet bases are particularly well suited to approximate (not necessarily periodic) smooth functions, which is the reason why we use here CDV wavelet bases. On the other hand, for example, the Fourier basis is only appropriate to approximate periodic functions and it is often not natural to assume that the structural function g_0 is periodic.

4. Theoretical analysis

4.1. Basic assumptions

We state some basic assumptions. We do not state here assumptions on priors, which will be stated in the theorems below. In what follows, let $C_1 > 1$ be some constant. First of all, we assume:

Assumption 1. (i) (X, W) has joint density $f_{X,W}(x, w)$ on $[0, 1]^2$ satisfying that $f_{X,W}(x, w) \leq C_1$, $\forall x, w \in [0, 1]$. (ii) Denote by $f_W(w)$ the density of W , i.e., $f_W(w) = \int f_{X,W}(x, w)dx$. Then, $f_W(w) \geq C_1^{-1}$, $\forall w \in [0, 1]$. (iii) $\sup_{w \in [0, 1]} \mathbb{E}[Y^2 | W = w] \leq C_1$.

Assumption 1 is a usual restriction in the literature, up to minor differences (see [Hall and Horowitz, 2005](#); [Horowitz, 2012](#)). Denote by $f_X(x)$ the density of X , i.e., $f_X(x) = \int f_{X,W}(x, w)dw$. Then, we have $f_X(x) \leq C_1$, $\forall x \in [0, 1]$ and $f_W(w) \leq C_1$, $\forall w \in [0, 1]$.

For identification of g_0 , we assume:

Assumption 2. The linear operator $K : L_2[0, 1] \rightarrow L_2[0, 1]$ is injective.

For smoothness of g_0 , as mentioned before, we assume:

Assumption 3. $\exists s > 1/2$, $g_0 \in \mathcal{G}^s$, where \mathcal{G}^s is either $B_{\infty, \infty}^s$ or $B_{2, 2}^s$.

We refer the reader to [Newey and Powell \(2003\)](#) and [d'Haultfoeuille \(2011\)](#) for discussion on the identification issue. We should note that restricting the domain of K to a “small” set, such as a Sobolev ball, would substantially relax Assumption 2, which however requires a different analysis. For the sake of simplicity, we assume the injectivity of K on the full domain.

Remark 6. [Liao and Jiang \(2011\)](#) formally allowed for the case in which (in our notation) K is not injective, i.e., $\{g : Kg = 0\} \neq \{0\}$. However, their Assumption 4.5 (i) indeed implies the injectivity of K when the basis used is an

orthonormal basis of $L_2[0, 1]$. In the case that K is not injective, their Assumption 4.5 requires us to have some *a priori* knowledge on the eigen-structure of K^*K (K^* denotes the adjoint of K), which is typically not available.

As discussed in Introduction, solving (2) is an ill-posed inverse problem. Thus, the statistical difficulty of estimating g_0 depends on the difficulty of “inverting” K , which is usually referred to as “ill-posedness” of the inverse problem (2). Typically, the ill-posedness is characterized by the decay rate of $\kappa_l \rightarrow 0$ (κ_l is the l -th largest singular value of K), which is plausible if K were known and the singular value decomposition of K were used (see Cavalier, 2008). However, here, K is unknown and the known wavelet basis $\{\phi_l\}$ is used instead of the singular value system. Thus, it is suitable to quantify the ill-posedness using the wavelet basis $\{\phi_l\}$. To this end, define

$$\tau_J = s_{\min}(\mathbb{E}[\phi^J(W)\phi^J(X)^T]) = s_{\min}((\langle \phi_l, K\phi_m \rangle)_{1 \leq l, m \leq 2^J}), \quad J \geq J_0.$$

This quantity corresponds to what is called “sieve measure of ill-posedness” in the literature (Blundell et al., 2007; Horowitz, 2012). We at least have to assume that $\tau_J > 0$ for all $J \geq J_0$. Note however that

$$\begin{aligned} \tau_J &= s_{\min}((\langle \phi_l, K\phi_m \rangle)_{1 \leq l, m \leq 2^J}) \\ &= \min_{g \in V_J, \|g\|=1} \|(\langle \phi_l, Kg \rangle)_{1 \leq l \leq 2^J}\|_{\ell^2} \\ &\leq \min_{g \in V_J, \|g\|=1} \|Kg\| \quad (\text{Bessel's inequality}) \\ &\leq \kappa_{2^J}, \quad (\text{Courant-Fischer-Weyl's minimax principle}) \end{aligned}$$

by which, necessarily, $\tau_J \rightarrow 0$ as $J \rightarrow \infty$. For this quantity, we assume:

Assumption 4. (i) $\exists r > 0$, $\tau_J \geq C_1^{-1}2^{-Jr}$, $\forall J \geq J_0$; (ii) $\|\mathbb{E}[\phi^J(W)(g_0 - P_J g_0)(X)]\|_{\ell^2} (= \|(\langle \phi_l, K(g_0 - P_J g_0) \rangle)_{l=1}^{2^J}\|_{\ell^2}) \leq C_1 \tau_J \|g_0 - P_J g_0\|$, $\forall J \geq J_0$.

Assumption 4 (i) lower bounds τ_J as $J \rightarrow \infty$, thereby quantifies the ill-posedness. Here we only consider mildly ill-posed cases for some technical reasons. This rules out e.g. the case in which the joint density $f_{X,W}(x, w)$ is analytic (see Kless, 1999, Theorem 15.20).

Assumption 4 (ii) is a “stability” condition about the bias $g_0 - P_J g_0$, which states that $K(g_0 - P_J g_0)$ is sufficiently “small” relative to $g_0 - P_J g_0$. Note that in the (ideal) case in which K is self-adjoint and $\{\phi_l\}$ is the eigen-basis of K , $\langle \phi_l, K(g_0 - P_J g_0) \rangle = 0$ for all $l = 1, \dots, 2^J$, in which case Assumption 4 (ii) is trivially satisfied. Assumption 4 (ii) allows more general situations in which K may not be self-adjoint and $\{\phi_l\}$ may not be the eigen-basis of K by allowing for a certain “slack”. This assumption, although looks technical, is common in the study of rates of convergence in estimation of the structural function g_0 . Indeed, essentially similar conditions have appeared in the past literature such as Blundell et al. (2007); Chen and Reiss (2011); Horowitz (2012). For example, Blundell et al. (2007, Assumption 6) essentially states (in our notation) that $\|K(g_0 - P_J g_0)\| \leq C_1 \tau_J \|g_0 - P_J g_0\|$, which implies our Assumption 4 (ii) since $\|(\langle \phi_l, K(g_0 - P_J g_0) \rangle)_{l=1}^{2^J}\|_{\ell^2} \leq \|K(g_0 - P_J g_0)\|$ (Bessel's inequality).

Remark 7. For given values of $C_1 > 1, M > 0, r > 0$ and $s > 1/2$, let $\mathcal{F} = \mathcal{F}(C_1, M, r, s)$ denote the set of all distributions of (Y, X, W) satisfying Assumptions 1-4 with $\|g_0\|_{s,\infty,\infty} \leq M$ in case of $\mathcal{G}^s = B_{\infty,\infty}^s$ and $\|g_0\|_{s,2,2} \leq M$ in case of $\mathcal{G}^s = B_{2,2}^s$. By Hall and Horowitz (2005); Chen and Reiss (2011), it is shown that the minimax rate of convergence (in $\|\cdot\|$) of estimation of g_0 over this distribution class \mathcal{F} is $n^{-s/(2r+2s+1)}$ as the sample size $n \rightarrow \infty$.

Suppose now that for some $\varepsilon_n \rightarrow 0$, $\sup_{F \in \mathcal{F}} \mathbb{E}_F[\Pi_n(g : \|g - g_0\| > \varepsilon_n \mid \mathcal{D}_n)] \rightarrow 0$. Then, by Theorem 2.5 of Ghosal et al. (2000), there exists a point estimator that converges (in probability) at least as fast as ε_n uniformly in $F \in \mathcal{F}$. Here, the quasi-posterior is not a proper posterior, but the proof of Ghosal et al. (2000, Theorem 2.5) applies to this case. By this, in the minimax sense, the fastest possible rate of contraction of the quasi-posterior distribution $\Pi_n(dg \mid \mathcal{D}_n)$ is $n^{-s/(2r+2s+1)}$.

4.2. Main results

In what follows, let $(Y_1, X_1, W_1), \dots, (Y_n, X_n, W_n)$ be i.i.d. observations of (Y, X, W) . Denote by $b_0^J = (b_{01}, \dots, b_{0,2^J})^T$ the vector of the first 2^J generalized Fourier coefficients of g_0 , i.e., $b_{0l} = \int \phi_l g_0$. Let $\|\cdot\|_{\text{TV}}$ denote the total variation norm between two distributions.

Theorem 1. Suppose that Assumptions 1-4 are satisfied. Take J_n in such a way that $J_n \rightarrow \infty$ and $2^{J_n} = o((n/\log n)^{1/(2r+1)})$. Let ϵ_n be a sequence of positive constants such that $\epsilon_n \rightarrow 0$ and $n\epsilon_n^2 \gtrsim 2^{J_n}$. Suppose that generating priors $\tilde{\Pi}_n$ has densities $\tilde{\pi}_n$ on $\mathbb{R}^{2^{J_n}}$ and satisfy the following conditions:

- P1) (Small ball condition) There exists a constant $C > 0$ such that for all n sufficiently large, $\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq \epsilon_n) \geq e^{-Cn\epsilon_n^2}$.
- P2) (Prior flatness condition) Let $\gamma_n = 2^{-J_n s} + 2^{J_n r} \epsilon_n$. There exists a sequence of constants $L_n \rightarrow \infty$ sufficiently slowly such that for all n sufficiently large, $\tilde{\pi}_n(b^{J_n})$ is positive for all $\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n$, and

$$\sup_{\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n, \|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n} \left| \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} - 1 \right| \rightarrow 0.$$

Then, for every sequence $M_n \rightarrow \infty$, we have

$$\tilde{\Pi}_n \left\{ b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > M_n (2^{-J_n s} + 2^{J_n r} \sqrt{2^{J_n}/n}) \mid \mathcal{D}_n \right\} \xrightarrow{P} 0. \quad (8)$$

Furthermore, assume that $2^{J_n} = o((n/\log n)^{1/(2r+3)})$. Then, we have

$$\begin{aligned} & \|\tilde{\Pi}_n\{b^{J_n} : \sqrt{n}(b^{J_n} - b_0^{J_n}) \in \cdot \mid \mathcal{D}_n\} \\ & \quad - N(\Delta_n, \Phi_{WX}^{-1} \Phi_{WW} \Phi_{XW}^{-1}(\cdot))\|_{\text{TV}} \xrightarrow{P} 0. \end{aligned} \quad (9)$$

Here, $\Delta_n = \sqrt{n} \Phi_{WX}^{-1} \mathbb{E}_n[\phi^{J_n}(W_i) R_i]$, $R_i = U_i + (g_0(X_i) - P_{J_n} g_0(X_i))$, $U_i = Y_i - g_0(X_i)$, $\Phi_{WX} = \mathbb{E}[\Phi^{J_n}(W) \Phi^{J_n}(X)^T]$, $\Phi_{XW} = \Phi_{WX}^T$, and $\Phi_{WW} = \mathbb{E}[\phi^{J_n}(W)^{\otimes 2}]$.

Proof. See Section 6. \square

First of all, since for $g = \sum_{l=1}^{2^{J_n}} b_l \phi_l$, $\|g - g_0\|^2 = \|g - P_J g_0\|^2 + \|g_0 - P_J g_0\|^2 \lesssim \|b^{J_n} - b_0^{J_n}\|_{\ell^2}^2 + 2^{-2J_n s}$, part (8) of Theorem 1 leads to that for every sequence $M_n \rightarrow \infty$,

$$\Pi_n \left\{ g : \|g - g_0\| > M_n(2^{-J_n s} + 2^{J_n r} \sqrt{2^{J_n}/n}) \mid \mathcal{D}_n \right\} \xrightarrow{P} 0,$$

which means that the rate of contraction of the quasi-posterior distribution $\Pi_n(dg \mid \mathcal{D}_n)$ is $\max\{2^{-J_n s}, 2^{J_n r} \sqrt{2^{J_n}/n}\}$.³ In many examples, for given $J_n \rightarrow \infty$ with $2^{J_n} = o((n/\log n)^{1/(2r+1)})$, condition P1) is satisfied with $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. Taking J_n in such a way that $2^{J_n} \sim n^{1/(2r+2s+1)}$, which leads to the optimal contraction rate, γ_n in condition P2) is $\sim n^{-s/(2r+2s+1)}(\log n)^{1/2}$. So condition P2) states that, to attain the optimal contraction rate, the prior density $\tilde{\pi}_n$ should be sufficiently “flat” in a ball with center $b_0^{J_n}$ and radius of order (essentially) $n^{-s/(2r+2s+1)}$. Some specific priors leading to the optimal contraction rate will be given in Section 5.

As noted before, in many examples, for given $J_n \rightarrow \infty$ with $2^{J_n} = o((n/\log n)^{1/(2r+1)})$, condition P1) is satisfied with $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. Inspection of the proof shows that, without condition P2), this already leads to contraction rate $\max\{2^{-J_n s}, 2^{J_n r} \sqrt{2^{J_n}(\log n)/n}\}$, which reduces to $(n/\log n)^{-s/(2r+2s+1)}$ by taking $2^{J_n} \sim (n/\log n)^{1/(2r+2s+1)}$. However, this rate is not fully satisfactory because of the appearance of the log term. Condition P2) is used to get rid of the log term.

Under a further integrability condition about U , $M_n \rightarrow \infty$ in (8) can be replaced by a large fixed constant M .

Theorem 2. *Suppose that all the conditions that guarantee (8) in Theorem 1 are satisfied. Furthermore, assume that $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \rightarrow 0$ as $\lambda \rightarrow \infty$. Then, there exists a constant $M > 0$ such that*

$$\tilde{\Pi}_n \left\{ b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > M(2^{-J_n s} + 2^{J_n r} \sqrt{2^{J_n}/n}) \mid \mathcal{D}_n \right\} \xrightarrow{P} 0. \quad (10)$$

Proof. See Appendix B. \square

The proof consists in establishing a concentration property of the random variable $\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2}$, which uses a truncation argument and Talagrand’s (1996) concentration inequality. A sufficient condition that guarantees that $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \rightarrow 0$ as $\lambda \rightarrow \infty$ is that $\exists \epsilon > 0$, $\sup_{w \in [0,1]} \mathbb{E}[|U|^{2+\epsilon} \mid W = w] < \infty$.

The second part of Theorem 1 states a Bernstein-von Mises type result for the quasi-posterior distribution $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$. A difference from the standard Bernstein-von Mises theorem is that the covariance matrix of the centering variable $\Phi_{WX}^{-1} \sqrt{n} \mathbb{E}_n[\phi^{J_n}(W_i)U_i]$ (without the bias part) is $\Phi_{WX}^{-1} \mathbb{E}[\sigma_0^2(W) \phi^{J_n}(W)^{\otimes 2}] \Phi_{XW}^{-1}$

³We have ignored the appearance of $M_n \rightarrow \infty$, which can be arbitrarily slow. A version in which M_n is replaced by a large fixed constant $M > 0$ is presented in Theorem 2.

with $\sigma_0^2(W) = \mathbb{E}[U^2 \mid W]$ and different from $\Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1}$ (which is the reason why we added “type”). This is a generic nature of quasi-posterior distributions. Even for finite dimensional models, generally, the covariance matrix of the centering variable does not coincide with that of the normal distribution approximating the quasi-posterior distribution (see Chernozhukov and Hong, 2003).

An alternative expression of (9) is stated as follows. Let \hat{b}^{J_n} denote a “maximum quasi-likelihood estimator” of $b_0^{J_n}$, i.e.,

$$\hat{b}^{J_n} \in \arg \max_{b^{J_n} \in \mathbb{R}^{2J_n}} p_{b^{J_n}}(\mathcal{D}_n).$$

Here, note that $\hat{g}(\cdot) := \phi^{J_n}(\cdot)^T \hat{b}^{J_n}$ is a maximum quasi-likelihood estimator of g_0 over V_{J_n} , i.e., $\hat{g} \in \arg \max_{g \in V_{J_n}} p_g(\mathcal{D}_n)$ and essentially the same as the sieve minimum distance estimator of Blundell et al. (2007). Under the assumptions of Theorem 1, with probability approaching one, $\hat{b}^{J_n} = \hat{\Phi}_{WX}^{-1} \mathbb{E}_n[\phi^{J_n}(W_i)Y_i] = b_0^{J_n} + \hat{\Phi}_{WX}^{-1} \mathbb{E}_n[\phi^{J_n}(W_i)R_i]$. Given the proof of Theorem 1, it is not hard to see that

$$\|\tilde{\Pi}_n(\cdot \mid \mathcal{D}_n) - N(\hat{b}^{J_n}, n^{-1}\Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1})(\cdot)\|_{TV} \xrightarrow{P} 0,$$

which is perhaps a more interpretable form of the asymptotic normality of the quasi-posterior distribution $\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n)$.

Finally, we consider the convergence rate of the quasi-Bayes estimator \hat{g}_{QB} of g_0 defined by (6).

Theorem 3. *Suppose that all the conditions of Theorem 2 are satisfied. Let \hat{g}_{QB} be the quasi-Bayes estimator defined by (6). Then, $\mathbb{P}\{\mathcal{D}_n : \int |g(x)|\Pi_n(dg \mid \mathcal{D}_n) < \infty, \forall x \in [0, 1]\} \rightarrow 1$, and there exists a constant $M > 0$ such that*

$$\begin{aligned} & \mathbb{P}\left[\|\hat{g}_{QB} - g_0\| \right. \\ & \quad \left. \leq M \max\{2^{-J_n s}, 2^{J_n r} \sqrt{2^{J_n}/n}, 2^{J_n r} \epsilon_n \varrho_n (\log n)^{1/2}\}\right] \rightarrow 1, \end{aligned} \quad (11)$$

where

$$\varrho_n := \sup_{\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n, \|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n} \left| \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} - 1 \right|.$$

Here ϵ_n, γ_n and L_n are given in the statement of Theorem 1.

Proof. See Section C. □

Theorem 3 is *not* directly deduced from Theorem 1. Indeed, $\|g - g_0\|$ may not be bounded on the support of Π_n since the support of Π_n is allowed to be unbounded in $\|\cdot\|$, and hence the argument used in Ghosal et al. (2000, p.506-p.507) can not apply here (in Ghosal et al. (2000), a typical distance to measure the goodness of a point estimator is the Hellinger distance and uniformly bounded). Hence, an additional work is needed to prove Theorem 3.

The convergence rate of the quasi-Bayes estimator is determined by the three terms: $2^{-J_n s}$, $2^{J_n r} \sqrt{2^{J_n}/n}$, and $2^{J_n r} \epsilon_n \varrho_n (\log n)^{1/2}$. The last term is typically small relative to the other two terms. Indeed, as noted before, in many examples, for given $J_n \rightarrow \infty$ with $2^{J_n} = o((n/\log n)^{1/(2r+1)})$, ϵ_n can be taken in such a way that $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. In that case $2^{J_n r} \epsilon_n \varrho_n (\log n)^{1/2} \sim 2^{J_n r} \sqrt{2^{J_n}/n} \times \varrho_n (\log n)$, and as long as $\varrho_n \rightarrow 0$ sufficiently fast, i.e., $\varrho_n = O((\log n)^{-1})$, the convergence rate of the quasi-Bayes estimator \hat{g}_{QB} reduces to $\max\{2^{-J_n s}, 2^{J_n r} \sqrt{2^{J_n}/n}\}$, which further reduces to $n^{-s/(2r+2s+1)}$ if we can take $2^{J_n} \sim n^{1/(2r+2s+1)}$. The rate $n^{-s/(2r+2s+1)}$ is minimax optimal under the present setting (see Remark 7). Note here that by inspection of the proof, $(\log n)^{1/2}$ in (11) indeed can be replaced by any other sequence slowly divergent as $n \rightarrow \infty$. We do not pursue the generality in this direction.

5. Prior specification: examples

In this section, we give some specific sieve priors for which the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, respectively). We consider two types of priors, namely, shrinking priors and non-shrinking priors. By a shrinking prior, we mean a prior that has smaller weights on b_l for larger l . A non-shrinking prior is a prior that is not a shrinking prior.⁴

5.1. Non-shrinking priors

We first consider non-shrinking priors.

Proposition 2. *Suppose that Assumptions 1-4 are satisfied. Consider the following two classes of prior distributions on $\mathbb{R}^{2^{J_n}}$:*

(Product prior) *Let $q(x)$ be a probability density function on \mathbb{R} such that for a constant $A > \sup_{l \geq 1} |b_{0l}|$: 1) $q(x)$ is positive on $[-A, A]$; 2) $\log q(x)$ is Lipschitz continuous on $[-A, A]$, i.e., there exists a constant $L > 0$ possibly depending on A such that $|\log q(x) - \log q(y)| \leq L|x - y|$, $\forall x, y \in [-A, A]$.*

Take the density of the generating prior by $\tilde{\pi}_n(b^{J_n}) = \prod_{l=1}^{2^{J_n}} q(b_l)$.

(Isotropic prior) *Let $r(x)$ be a probability density function on $[0, \infty)$ having all moments such that: 1) for a constant $A > \|g_0\|$, $r(x)$ is positive and continuous on $[0, A]$; 2) for a constant $c > 0$, $\int_0^\infty x^{k-1} r(x) dx \leq e^{ck \log k}$ for all k sufficiently large. Take the density of the generating prior by $\tilde{\pi}_n(b^{J_n}) \propto r(\|b^{J_n}\|_{\ell_2})$.*

Let $2^{J_n} \sim n^{1/(2r+2s+1)}$. Then, in either case, for every sequence $M_n \rightarrow \infty$, we have $\Pi_n\{g : \|g - g_0\| > M_n n^{-s/(2s+2r+1)} \mid \mathcal{D}_n\} \xrightarrow{P} 0$. Furthermore, if $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \rightarrow 0$ as $\lambda \rightarrow \infty$, then there exists a constant $M > 0$ such that $\Pi_n\{g : \|g - g_0\| > M n^{-s/(2s+2r+1)} \mid \mathcal{D}_n\} \xrightarrow{P} 0$.

⁴This terminology is only for convenience and not strictly well-defined.

Proof. See Appendix D. \square

Proposition 2 shows that a wide class of non-shrinking priors lead to the optimal contraction rate. In either case of product or isotropic priors, the constant A is not necessarily known, which allows $q(x)$ and $r(x)$ to have unbounded support. For example, in the former case, $q(x)$ may be the density of the standard normal distribution, in which case A can be taken to be arbitrarily large. Likewise, in the latter case, $r(x)$ may be the density of an exponential distribution: $r(x) = \lambda e^{-\lambda x}$, $x \geq 0$ for some $\lambda > 0$. In the isotropic prior case, $r(x)$ should have all moments, i.e., $\int_0^\infty x^k r(x) dx < \infty$ for all $k \geq 1$, which ensures that $\tilde{\pi}_n(b^{J_n}) \propto r(\|b^{J_n}\|_{\ell^2})$ is a proper distribution on $\mathbb{R}^{2^{J_n}}$ for every $n \geq 1$.

For the quasi-Bayes estimator \hat{g}_{QB} , we have:

Proposition 3. *Suppose that Assumptions 1-4 are satisfied. Furthermore, assume that $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \rightarrow 0$ as $\lambda \rightarrow \infty$. Consider the two classes of prior distributions on $\mathbb{R}^{2^{J_n}}$ given in Proposition 2. In the isotropic prior case, assume further that $r(x)$ is Lipschitz continuous on $[0, A]$. Let $2^{J_n} \sim n^{1/(2r+2s+1)}$. Then, in either case of product or isotropic priors, there exists a constant $M > 0$ such that $\mathbb{P}\{\|\hat{g}_{QB} - g_0\| > Mn^{-s/(2r+2s+1)}\} \rightarrow 0$.*

Proof. See Appendix D. \square

5.2. Shrinking priors

We next consider shrinking priors.

Proposition 4. *Suppose that Assumptions 1-4 are satisfied. Furthermore, assume that $\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) \mid W = w] \rightarrow 0$ as $\lambda \rightarrow \infty$. Consider either case (a) or case (b) below:*

Case (a) $g_0 \in B_{\infty,\infty}^s$, and let the generating prior $\tilde{\Pi}_n$ be the distribution of $b^{J_n} = (b_1, \dots, b_{2^{J_n}})^T$ constructed by the following steps: 1) Generate $u_1, \dots, u_{2^{J_n}} \sim U[-A_n, A_n]$ i.i.d. with $A_n \sim (\log n)\sqrt{2^{J_n}}$; 2) Let $b_l = u_l$ for $l = 1, \dots, 2^{J_0}$ and $b_{2^j+k} = 2^{-j(s+1/2)}u_{2^j+k}$ for $k = 1, \dots, 2^j$; $j = J_0, \dots, J_n - 1$.

Case (b) $g_0 \in B_{2,2}^s$, and let the generating prior $\tilde{\Pi}_n$ be the distribution of $b^{J_n} = (b_1, \dots, b_{2^{J_n}})^T$ constructed by the following steps: 1) Generate $u_1, \dots, u_{2^{J_n}} \sim N(0, A_n^2)$ i.i.d. with $A_n \sim (\log n)\sqrt{2^{J_n}}$; 2) Let $b_l = u_l$ for $l = 1, \dots, 2^{J_0}$ and $b_{2^j+k} = 2^{-j(s+1/2)}u_{2^j+k}$ for $k = 1, \dots, 2^j$; $j = J_0, \dots, J_n - 1$.

Let $2^{J_n} \sim n^{1/(2r+2s+1)}$. Then, in either case, there exists a constant $M > 0$ such that $\Pi_n\{g : \|g - g_0\| > Mn^{-s/(2s+2r+1)} \mid \mathcal{D}_n\} \xrightarrow{P} 0$ and $\mathbb{P}\{\|\hat{g}_{QB} - g_0\| > Mn^{-s/(2r+2s+1)}\} \rightarrow 0$.

Proof. See Appendix D. \square

Proposition 4 shows that a class of shrinking priors, suitably rescaled by the factor $A_n \rightarrow \infty$, leads to the optimal convergence rate. The rescaling is used to guarantee sufficient “flatness” of the priors.

From a theoretical point of view, using non-shrinking priors is sufficient to achieve the optimal convergence rate. However, practically, it would be beneficial to use shrinking priors since e.g. putting the prior in case (a) roughly means adding a penalty on the magnitude of the Hölder(-Zygmund) norm, which would result in a numerical stability (likewise, putting the prior in case (b) roughly means adding a penalty on the magnitude of the Sobolov norm).

6. Proof of Theorem 1

Before proving Theorem 1, we first prepare some technical lemmas (Lemmas 1-4) and establish preliminary rates of contraction for the quasi-posterior distribution (Proposition 5). Proofs of Lemmas 1, 2 and 4 are given in Appendix A. For the notational convenience, define the matrices

$$\hat{\Phi}_{WX} = \mathbb{E}_n[\phi^{J_n}(W_i)\phi^{J_n}(X_i)^T], \quad \hat{\Phi}_{XW} = \hat{\Phi}_{WX}^T, \quad \text{and} \quad \hat{\Phi}_{WW} = \mathbb{E}_n[\phi^{J_n}(W_i)^{\otimes 2}].$$

Recall that $\Phi_{WX} = \mathbb{E}[\hat{\Phi}_{WX}] = \mathbb{E}[\phi^{J_n}(W)\phi^{J_n}(X)^T]$ and $\Phi_{WW} = \mathbb{E}[\hat{\Phi}_{WW}] = \mathbb{E}[\phi^{J_n}(W)^{\otimes 2}]$.

Lemma 1. *Suppose that Assumptions 1-4 are satisfied. Let $J_n \rightarrow \infty$ as $n \rightarrow \infty$. (i) There exists a constant $D > 0$ such that $\sup_{w \in [0,1]} \|\phi^J(w)\|_{\ell^2} \leq D2^{J/2}$ for all $J \geq J_0$. (ii) $C_1^{-1} \leq s_{\min}(\mathbb{E}[\phi^J(W)^{\otimes 2}]) \leq s_{\max}(\mathbb{E}[\phi^J(W)^{\otimes 2}]) \leq C_1$ and $s_{\max}(\mathbb{E}[\phi^J(W)\phi^J(X)^T]) \leq C_1$ for all $J \geq J_0$. (iii) If $J_n 2^{J_n}/n \rightarrow 0$, $\|\hat{\Phi}_{WW} - \Phi_{WW}\|_{\text{op}} = O_P(\sqrt{J_n 2^{J_n}/n})$ and $\|\hat{\Phi}_{WX} - \Phi_{WX}\|_{\text{op}} = O_P(\sqrt{J_n 2^{J_n}/n})$. (iv) $\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 = O_P(2^{J_n}/n + \tau_{J_n}^2 2^{-2J_n s})$. (v) If $J_n 2^{J_n(2r+1)}/n \rightarrow 0$, $s_{\min}(\hat{\Phi}_{WX}) \geq (1 - o_P(1))\tau_{J_n}$.*

The following lemma characterizes the total variation convergence between two centered multivariate distributions with increasing dimensions via the speed of convergence between the corresponding covariance matrices.

Lemma 2. *Let Σ_n be a sequence of symmetric positive definite matrices of dimension $k_n \rightarrow \infty$ as $n \rightarrow \infty$ such that $\|\Sigma_n - I_{k_n}\|_{\text{op}} = o(k_n^{-1})$. Then, as $n \rightarrow \infty$,*

$$\int |dN(0, \Sigma_n)(x) - dN(0, I_{k_n})(x)| dx \rightarrow 0.$$

The following lemma is due to Lemma 4 of Bontemps (2011).

Lemma 3. *Let Z be a k -vector of constants with $k \in \mathbb{N}$. Then, $\|N(Z, I_k) - N(0, I_k)\|_{\text{TV}} \leq \|Z\|_{\ell^2}/\sqrt{2\pi}$.*

The following lemma was used in the proof of Theorem 1.

Lemma 4. *Let \hat{A}_n be a sequence of random $k_n \times k_n$ matrices where k_n is either bounded or $k_n \rightarrow \infty$ as $n \rightarrow \infty$. Suppose that there exist sequences of positive constants ϵ_n, δ_n and a sequence of non-random, non-singular $k_n \times k_n$ matrices A_n such that $\epsilon_n \rightarrow 0, \delta_n \rightarrow 0, s_{\min}(A_n) \gtrsim \epsilon_n, \|\hat{A}_n - A_n\|_{\text{op}} = O_P(\delta_n)$ and $\epsilon_n^{-1}\delta_n \rightarrow 0$. Then, \hat{A}_n is non-singular with probability approaching one and $\|\hat{A}_n^{-1}A_n - I_{k_n}\|_{\text{op}} = O_P(\epsilon_n^{-1}\delta_n)$. Likewise, $\|A_n\hat{A}_n^{-1} - I_{k_n}\|_{\text{op}} = O_P(\epsilon_n^{-1}\delta_n)$.*

The following proposition gives preliminary rates of contraction for the quasi-posterior distribution.

Proposition 5 (Preliminary contraction rates). *Suppose that Assumptions 1-4 are satisfied. Take J_n in such a way that $J_n \rightarrow \infty$ and $2^{J_n} = o((n/\log n)^{1/(2r+1)})$. Let ϵ_n be a sequence of positive constants such that $\epsilon_n \rightarrow 0$ and $\sqrt{n}\epsilon_n \rightarrow \infty$. Assume that a sequence of generating priors $\tilde{\Pi}_n$ satisfies condition P1) of Theorem 1. Define the data-dependent, empirical seminorm $\|\cdot\|_{\mathcal{D}_n}$ on $\mathbb{R}^{2^{J_n}}$ by*

$$\|b^{J_n}\|_{\mathcal{D}_n} = \|\hat{\Phi}_{WX}b^{J_n}\|_{\ell^2}, \quad b^{J_n} \in \mathbb{R}^{2^{J_n}}.$$

Then, we have for every sequence $M_n \rightarrow \infty$,

$$\tilde{\Pi}_n\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n(\epsilon_n + \tau_{J_n}2^{-J_n s}) \mid \mathcal{D}_n\} \xrightarrow{P} 0.$$

Proof of Proposition 5. Let $\delta_n = \epsilon_n + \tau_{J_n}2^{-J_n s}$. We wish to show that there exists a constant $c_0 > 0$ such that

$$\mathbb{P}\left\{\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n) \leq e^{-c_0 M_n^2 n \delta_n^2}\right\} \rightarrow 1. \quad (12)$$

Note that since $\sqrt{n}\epsilon_n \rightarrow \infty$, $n\delta_n^2 \geq n\epsilon_n^2 \rightarrow \infty$. Below, c_1, c_2, \dots are some positive constants of which the values are understood in the context.

Recall $R_i = U_i + \sum_{l=2^{J_n}+1}^\infty b_{0l}\phi_l(X_i) = U_i + (g_0(X_i) - P_{J_n}g_0(X_i))$. Then, for $b^{J_n} \in \mathbb{R}^{2^{J_n}}$,

$$\begin{aligned} \mathbb{E}_n[\hat{m}^2(W_i, b^{J_n})] &= -2(b^{J_n} - b_0^{J_n})^T \hat{\Phi}_{XW} \hat{\Phi}_{WW}^{-1} \mathbb{E}_n[\phi^{J_n}(W_i)R_i] \\ &\quad + (b^{J_n} - b_0^{J_n})^T \hat{\Phi}_{XW} \hat{\Phi}_{WW}^{-1} \hat{\Phi}_{WX} (b^{J_n} - b_0^{J_n}) \\ &\quad + \mathbb{E}_n[\phi^{J_n}(W_i)R_i]^T \hat{\Phi}_{WW}^{-1} \mathbb{E}_n[\phi^{J_n}(W_i)R_i]. \end{aligned} \quad (13)$$

Since the last term is independent of b^{J_n} , it is canceled out in the quasi-posterior distribution. Denote by $\ell_{b^{J_n}}(\mathcal{D}_n)$ the sum of the first two terms in (13). Then,

$$\tilde{\Pi}_n(db^{J_n} \mid \mathcal{D}_n) \propto \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\} \tilde{\Pi}_n(db^{J_n}).$$

Using the fact that for any $x, y, c \in \mathbb{R}$ with $c > 0$, $2xy \leq cx^2 + c^{-1}y^2$, we have

$$\begin{aligned} \ell_{b^{J_n}}(\mathcal{D}_n) &\geq (\hat{\lambda}_{\min} - c)\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n}^2 \\ &\quad - c^{-1}\hat{\lambda}_{\max}^2\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2, \quad \forall c > 0, \end{aligned} \quad (14)$$

where $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{\max}$ are the minimum and maximum eigenvalues of the matrix $\hat{\Phi}_{WW}^{-1}$, respectively. Likewise, we have

$$\begin{aligned} \ell_{b^{J_n}}(\mathcal{D}_n) &\leq (\hat{\lambda}_{\max} + c)\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n}^2 \\ &\quad + c^{-1}\hat{\lambda}_{\max}^2\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2, \quad \forall c > 0. \end{aligned} \quad (15)$$

Define the event

$$\mathcal{E}_{1n} = \{\mathcal{D}_n : \hat{\lambda}_{\min} < 0.5C_1^{-1}\} \cup \{\mathcal{D}_n : \hat{\lambda}_{\max} > 1.5C_1\} \\ \cup \{\mathcal{D}_n : \|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 > M_n\delta_n^2\}.$$

Construct the “tests” ω_n by $\omega_n = 1(\mathcal{E}_{1n})$. Then, we have

$$\begin{aligned} & \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n) \\ &= \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n)\{\omega_n + (1 - \omega_n)\} \\ &\leq \omega_n + \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n)(1 - \omega_n). \end{aligned} \quad (16)$$

By Lemmas 1 (ii)-(iv), we have $\mathbb{P}(\omega_n = 1) = \mathbb{P}(\mathcal{E}_{1n}) \rightarrow 0$.

For the second term in (16), taking $c > 0$ sufficiently small in (14), we have

$$\begin{aligned} & (1 - \omega_n) \int_{\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n} \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\} \tilde{\Pi}_n(db^{J_n}) \\ &\leq \exp\{-c_1 M_n^2 n \delta_n^2 + O(M_n n \delta_n^2)\} \\ &\leq e^{-c_2 M_n^2 n \delta_n^2}. \end{aligned}$$

On the other hand, taking, say $c = 1$ in (15), we have

$$\begin{aligned} & (1 - \omega_n) \int \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\} \tilde{\Pi}_n(db^{J_n}) \\ &\geq (1 - \omega_n) \int_{\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} \leq \sqrt{M_n}\epsilon_n} \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\} \tilde{\Pi}_n(db^{J_n}) \\ &\geq (1 - \omega_n) e^{-c_3 M_n n \epsilon_n^2} \int_{\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} \leq \sqrt{M_n}\epsilon_n} \tilde{\Pi}_n(db^{J_n}). \end{aligned}$$

Denote by \hat{s}_{\max} the maximum singular value of the matrix $\hat{\Phi}_{WX}$, so that

$$\|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} \leq \hat{s}_{\max} \|b^{J_n} - b_0^{J_n}\|_{\ell^2}.$$

Define the event $\mathcal{E}_{2n} = \{\mathcal{D}_n : \hat{s}_{\max} \leq 1.5C_1\}$. By Lemmas 1 (ii) and (iii), we have $\mathbb{P}(\mathcal{E}_{2n}) \rightarrow 1$. Since $M_n \rightarrow \infty$, for all n sufficiently large, we have

$$\begin{aligned} & 1(\mathcal{E}_{2n})(1 - \omega_n) \int \exp\{-(n/2)\ell_{b^{J_n}}(\mathcal{D}_n)\} \tilde{\Pi}_n(db^{J_n}) \\ &\geq 1(\mathcal{E}_{2n})(1 - \omega_n) e^{-c_3 M_n n \epsilon_n^2} \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq \epsilon_n) \\ &\geq 1(\mathcal{E}_{2n})(1 - \omega_n) e^{-c_3 M_n n \epsilon_n^2 - C n \epsilon_n^2} \\ &\geq 1(\mathcal{E}_{2n})(1 - \omega_n) e^{-c_4 M_n n \epsilon_n^2}, \end{aligned}$$

where the second inequality is due to the small ball condition P1). Summarizing, we have

$$\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\mathcal{D}_n} > M_n\delta_n \mid \mathcal{D}_n)(1 - \omega_n) \leq 1(\mathcal{E}_{2n}^c) + e^{-c_2 M_n^2 n \delta_n^2 + c_4 M_n n \epsilon_n^2}.$$

Since $\epsilon_n \leq \delta_n$, we obtain (12) for a sufficiently small $c_0 > 0$. \square

We are now in position to prove Theorem 1. We will say that a sequence of random variables A_n is *eventually* bounded by another sequence of random variables B_n if $\mathbb{P}(A_n \leq B_n) \rightarrow 1$ as $n \rightarrow \infty$.

Proof of Theorem 1. We first note that by Lemmas 1 (ii), (iii) and (v), the matrices $\hat{\Phi}_{WX}$ and $\hat{\Phi}_{WW}$ are non-singular with probability approaching one. Conditional on \mathcal{D}_n , define the rescaled “parameter” $\theta^{J_n} = (\theta_1, \dots, \theta_{2^{J_n}})^T = \sqrt{n}\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})$. By (13), the corresponding “quasi-posterior” density for θ^{J_n} is given by

$$\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) d\theta^{J_n} \propto \tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1} \theta^{J_n} / \sqrt{n}) dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n}) d\theta^{J_n},$$

where $\tilde{\Delta}_n = \sqrt{n}\mathbb{E}_n[\phi^{J_n}(W_i)R_i]$ (this operation is valid as soon as $\hat{\Phi}_{WX}$ and $\hat{\Phi}_{WW}$ are non-singular, of which the probability is approaching one).

The proof of Theorem 1 consists of 3 steps. After Step 1, we will turn to the proof of (8). The remaining two steps are devoted to the proof of (9).

Step 1. We first show that

$$\int |\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})| d\theta^{J_n} \xrightarrow{P} 0. \quad (17)$$

In this step, we do not assume $2^{J_n} = o((n/\log n)^{1/(2r+3)})$. As before, let $\delta_n = \epsilon_n + \tau_{J_n} 2^{-J_n s}$. By Proposition 5, for every sequence $M_n \rightarrow \infty$,

$$\int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} \pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) d\theta^{J_n} = 1 + o_P(1),$$

by which we have

$$\begin{aligned} & \text{Left side of (17)} \\ & \leq \int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} |\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})| d\theta^{J_n} \\ & \quad + \int_{\|\theta^{J_n}\|_{\ell^2} > M_n \sqrt{n} \delta_n} dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n}) d\theta^{J_n} + o_P(1). \end{aligned} \quad (18)$$

By Lemma 1 (iv), $\|\tilde{\Delta}_n\|_{\ell^2} = O_P(\sqrt{n} \delta_n)$, and by Lemmas 1 (ii) and (iii), $(1 - o_P(1))C_1^{-1} \leq s_{\min}(\hat{\Phi}_{WW}) \leq s_{\max}(\hat{\Phi}_{WW}) \leq (1 + o_P(1))C_1$, so that the second integral is eventually bounded by

$$\int_{\|\theta^{J_n}\|_{\ell^2} > \sqrt{M_n n} \delta_n} dN(0, I_{2^{J_n}})(\theta^{J_n}) d\theta^{J_n}. \quad (19)$$

Here, note that M_n is replaced by $\sqrt{M_n}$ to “absorb” the constant. By Borell’s inequality for Gaussian measures (see, for example, [van der Vaart and Wellner, 1996](#), Lemma A.2.2), for all $x > 0$,

$$\mathbb{P}(\|N(0, I_{2^{J_n}})\|_{\ell^2} > \sqrt{2^{J_n}} + x) \leq 2^{-x^2/2}. \quad (20)$$

Here since $n\delta_n^2 \geq n\epsilon_n^2 \gtrsim 2^{J_n}$, $\sqrt{M_n n}\delta_n/\sqrt{2^{J_n}} \rightarrow \infty$, so that the integral in (19) is $o(1)$.

It remains to show that the first integral in (18) is $o_P(1)$. This step uses a standard cancellation argument. Let $\mathcal{C}_n := \{\theta^{J_n} \in \mathbb{R}^{2^{J_n}} : \|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n}\delta_n\}$. First, provided that $\|\hat{\Phi}_{WX}^{-1}\|_{\text{op}} \leq 1.5\tau_{J_n}^{-1}$, for all $\theta^{J_n} \in \mathcal{C}_n$, $\|\hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n}\|_{\ell^2} \leq 1.5M_n\tau_{J_n}^{-1}\delta_n \leq 1.5M_n(2^{-J_n s} + C_1 2^{J_n r}\epsilon_n) \sim M_n\gamma_n$. So taking $M_n \rightarrow \infty$ sufficiently slowly such that $M_n = o(L_n)$, $\|\hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n}\|_{\ell^2} \leq L_n\gamma_n$ and hence $\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n}) > 0$ for all n sufficiently large. Here, by Lemma 1 (v), we have $\mathbb{P}(\|\hat{\Phi}_{WX}^{-1}\|_{\text{op}} \leq 1.5\tau_{J_n}^{-1}) \rightarrow 1$.

Suppose that $\|\hat{\Phi}_{WX}^{-1}\|_{\text{op}} \leq 1.5\tau_{J_n}^{-1}$. Let $\pi_{n,\mathcal{C}_n}^*(\theta^{J_n} \mid \mathcal{D}_n)$ and $dN^{\mathcal{C}_n}(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ denote the probability densities obtained by first restricting $\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)$ and $dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ to the ball \mathcal{C}_n and then renormalizing, respectively. By the first part of the present proof, replacing $\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)$ and $dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ by $\pi_{n,\mathcal{C}_n}^*(\theta^{J_n} \mid \mathcal{D}_n)$ and $dN^{\mathcal{C}_n}(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ respectively in the first integral in (18) has impact at most $o_P(1)$. Then, abbreviating $\pi_{n,\mathcal{C}_n}^*(d\theta^{J_n} \mid \mathcal{D}_n)$ by π_{n,\mathcal{C}_n}^* , $dN^{\mathcal{C}_n}(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ by $dN^{\mathcal{C}_n}$, $dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ by dN , and $\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n})$ by $\tilde{\pi}_n$, we have

$$\begin{aligned} \int |\pi_{n,\mathcal{C}_n}^* - dN^{\mathcal{C}_n}| &= \int \left| 1 - \frac{dN^{\mathcal{C}_n}}{\pi_{n,\mathcal{C}_n}^*} \right| \pi_{n,\mathcal{C}_n}^* \\ &= \int \left| 1 - \frac{dN/\int_{\mathcal{C}_n} dN}{\tilde{\pi}_n dN/\int_{\mathcal{C}_n} \tilde{\pi}_n dN} \right| \pi_{n,\mathcal{C}_n}^* \\ &= \int \left| 1 - \frac{\int_{\mathcal{C}_n} \tilde{\pi}_n dN}{\tilde{\pi}_n \int_{\mathcal{C}_n} dN} \right| \pi_{n,\mathcal{C}_n}^* \\ &= \int \left| 1 - \frac{\int_{\mathcal{C}_n} \tilde{\pi}_n dN^{\mathcal{C}_n}}{\tilde{\pi}_n} \right| \pi_{n,\mathcal{C}_n}^*. \end{aligned}$$

By the convexity of the map $x \mapsto |1 - x|$ and Jensen's inequality, the last expression is bounded by

$$\sup_{\theta^{J_n} \in \mathcal{C}_n, \tilde{\theta}^{J_n} \in \mathcal{C}_n} \left| 1 - \frac{\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\theta^{J_n}/\sqrt{n})}{\tilde{\pi}_n(b_0^{J_n} + \hat{\Phi}_{WX}^{-1}\tilde{\theta}^{J_n}/\sqrt{n})} \right|,$$

which is eventually bounded by

$$\sup_{\|b^{J_n}\|_{\ell^2} \leq L_n\gamma_n, \|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n\gamma_n} \left| 1 - \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} \right|.$$

The last expression goes to zeros as $n \rightarrow \infty$ by the prior flatness condition P2).

We now turn to the proof of (8). Take any $M_n \rightarrow \infty$ (this M_n may be different

from the previous M_n). By Step 1, we have

$$\begin{aligned} & \sup_{z>0} \left| \tilde{\Pi}_n \{b^{J_n} : \|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} > z \mid \mathcal{D}_n\} \right. \\ & \quad \left. - \int_{\|\theta^{J_n}\|_{\ell^2} > z} dN(n^{-1/2}\tilde{\Delta}_n, n^{-1}\hat{\Phi}_{WW})(\theta^{J_n})d\theta^{J_n} \right| \xrightarrow{P} 0. \end{aligned}$$

Here, by Lemma 1 (v), we have

$$\begin{aligned} \|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} & \geq s_{\min}(\hat{\Phi}_{WX})\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \\ & \geq (1 - o_P(1))\tau_{J_n}\|b^{J_n} - b_0^{J_n}\|_{\ell^2}, \end{aligned}$$

by which we have, uniformly in $z > 0$,

$$\begin{aligned} & \tilde{\Pi}_n \{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > 2\tau_{J_n}^{-1}z \mid \mathcal{D}_n\} \\ & \leq \tilde{\Pi}_n \{b^{J_n} : \|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} > z \mid \mathcal{D}_n\} + o_P(1) \\ & \leq \int_{\|\theta^{J_n}\|_{\ell^2} > z} dN(n^{-1/2}\tilde{\Delta}_n, n^{-1}\hat{\Phi}_{WW})(\theta^{J_n})d\theta^{J_n} + o_P(1). \end{aligned}$$

By Markov's inequality, the integral in the last expression is bounded by

$$\frac{1}{nz^2} \{ \|\tilde{\Delta}_n\|_{\ell^2}^2 + \text{tr}(\hat{\Phi}_{WW}) \}.$$

By Lemmas 1 (ii)-(iv), we have $\|\tilde{\Delta}_n\|_{\ell^2}^2 + \text{tr}(\hat{\Phi}_{WW}) = O_P(2^{J_n} + n\tau_{J_n}^2 2^{-2J_n s})$. Therefore, we conclude that, taking $z = M_n(\tau_{J_n} 2^{-J_n s} + \sqrt{2^{J_n}/n})$, $\tilde{\Pi}_n \{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} > 2M_n(2^{-J_n s} + \tau_{J_n}^{-1}\sqrt{2^{J_n}/n}) \mid \mathcal{D}_n\} \xrightarrow{P} 0$, which leads to the contraction rate result (8).

In what follows, we assume $2^{J_n} = o((n/\log n)^{1/(2r+3)})$, and prove the asymptotic normality result (9).

Step 2. (Replacement of $\hat{\Phi}_{WW}$ by Φ_{WW}). This step shows that

$$\int |dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n}) - dN(\tilde{\Delta}_n, \Phi_{WW})(\theta^{J_n})| d\theta^{J_n} \xrightarrow{P} 0,$$

which is equivalent to

$$\int |dN(0, \hat{\Phi}_{WW})(\theta^{J_n}) - dN(0, \Phi_{WW})(\theta^{J_n})| d\theta^{J_n} \xrightarrow{P} 0.$$

By Lemmas 1 (ii), (iii) and Lemma 2, this follows if $\sqrt{J_n 2^{J_n}/n} = o(2^{-J_n})$, i.e., $J_n 2^{3J_n} = o(n)$, which is satisfied since $2^{J_n} = o((n/\log n)^{1/(2r+3)})$.

Step 3. (Replacement of $\hat{\Phi}_{WX}$ by Φ_{WX}). We have shown that

$$\int |\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\tilde{\Delta}_n, \Phi_{WW})(\theta^{J_n})| d\theta^{J_n} \xrightarrow{P} 0.$$

By Scheffé's lemma, this means that

$$\|\tilde{\Pi}_n\{b^{J_n} : \sqrt{n}\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n}) \in \cdot \mid \mathcal{D}_n\} - N(\tilde{\Delta}_n, \Phi_{WW})(\cdot)\|_{TV} \xrightarrow{P} 0,$$

or,

$$\|\tilde{\Pi}_n\{b^{J_n} : \sqrt{n}(b^{J_n} - b_0^{J_n}) \in \cdot \mid \mathcal{D}_n\} - N(\hat{\Phi}_{WX}^{-1}\tilde{\Delta}_n, \hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1})(\cdot)\|_{TV} \xrightarrow{P} 0.$$

The last expression is asymptotically valid since $\hat{\Phi}_{WX}$ is non-singular with probability approaching one. The remaining step is to replace $\hat{\Phi}_{WX}$ by Φ_{WX} . This step requires a special care since the minimum singular value of Φ_{WX} (while positive) is approaching zero as $n \rightarrow \infty$. To conclude the theorem, it suffices to show the following two assertions:

$$\begin{aligned} & \|N(\hat{\Phi}_{WX}^{-1}\tilde{\Delta}_n, \hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1}) \\ & \quad - N(\hat{\Phi}_{WX}^{-1}\tilde{\Delta}_n, \hat{\Phi}_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1})\|_{TV} \xrightarrow{P} 0, \end{aligned} \quad (21)$$

$$\begin{aligned} & \|N(\hat{\Phi}_{WX}^{-1}\tilde{\Delta}_n, \Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1}) \\ & \quad - N(\Phi_{WX}^{-1}\tilde{\Delta}_n, \Phi_{WX}^{-1}\Phi_{WW}\Phi_{XW}^{-1})\|_{TV} \xrightarrow{P} 0. \end{aligned} \quad (22)$$

Note that $\Phi_{WX}^{-1}\tilde{\Delta}_n = \Delta_n$.

Proof of (21): Assertion (21) reduces to

$$\|N(0, \Phi_{WX}\hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1}\Phi_{XW}) - N(0, \Phi_{WW})\|_{TV} \xrightarrow{P} 0.$$

By Lemmas 1 (ii), (iii) and Lemma 3, $\|\Phi_{WX}\hat{\Phi}_{WX}^{-1}\Phi_{WW}\hat{\Phi}_{XW}^{-1}\Phi_{XW} - \Phi_{WW}\|_{\text{op}} = O_P(2^{J_n r} \sqrt{J_n 2^{J_n}/n}) = o_P(2^{-J_n})$ (the last equality follows by $2^{J_n} = o((n/\log n)^{1/(2r+3)})$). Since $C_1^{-1} \leq s_{\min}(\Phi_{WW}) \leq s_{\max}(\Phi_{WW}) \leq C_1$, the desired conclusion follows from Lemma 2.

Proof of (22): Assertion (22) reduces to

$$\|N((\Phi_{WX}\hat{\Phi}_{WX}^{-1} - I_{2^{J_n}})\tilde{\Delta}_n, \Phi_{WW}) - N(0, \Phi_{WW})\|_{TV} \xrightarrow{P} 0.$$

By Lemma 3, and the fact that $s_{\min}(\Phi_{WW}) \geq C_1^{-1}$, the left side is $\lesssim \|(\Phi_{WX}\hat{\Phi}_{WX}^{-1} - I_{2^{J_n}})\tilde{\Delta}_n\|_{\ell^2}$. Here, we have

$$\begin{aligned} \|(\Phi_{WX}\hat{\Phi}_{WX}^{-1} - I_{2^{J_n}})\tilde{\Delta}_n\|_{\ell^2} & \leq \|\Phi_{WX}\hat{\Phi}_{WX}^{-1} - I_{k_n}\|_{\text{op}}\|\tilde{\Delta}_n\|_{\ell^2} \\ & = O_P(\tau_{J_n}^{-1} \sqrt{J_n 2^{J_n}/n}) \times O_P(\sqrt{n}\tau_{J_n} 2^{-J_n s} + \sqrt{2^{J_n}}) \\ & = o_P(1), \end{aligned}$$

where the second line is due to Lemmas 1 (iii), (iv) and Lemma 3. The last line follows from $s > 1/2$ and $2^{J_n} = o((n/\log n)^{1/(2r+3)})$.

Steps 1-3 lead to the asymptotic normality result (9). \square

7. Discussion

In this paper, we have studied the asymptotic properties of quasi-posterior distributions against sieve priors in the NPIV model and given some specific priors for which the quasi-posterior distribution (the quasi-Bayes estimator) attains the minimax optimal rate of contraction (convergence, respectively). These results greatly sharpen the previous work of [Liao and Jiang \(2011\)](#).

We end the paper with some remarks on the direction of future work. First, as also noted by [Liao and Jiang \(2011\)](#), (adaptive) selection of the resolution level J_n in a (quasi-)Bayesian or “empirical” Bayesian approach is an important topic to be investigated. Second, a (quasi-)Bayesian analysis is typically useful in the analysis of complex models in which frequentistic estimation is difficult to implement due to non-differentiability/non-convex nature of loss functions. This usefulness comes from the fact that a (quasi-)Bayesian approach is typically able to avoid numerical optimization. See [Chernozhukov and Hong \(2003\)](#) and [Liu et al. \(2007\)](#) for the finite dimensional case. In infinite dimensional models, such a computational challenge in frequentistic estimation occurs in the analysis of nonparametric instrumental quantile regression models ([Horowitz and Lee, 2007](#); [Chen and Pouzo, 2011](#); [Gagliardini and Scaillet, 2011](#)). In that model, a typical loss function contains the indicator function and hence highly non-convex. In such a case, the computation of an optimal solution is by itself difficult, and a solution obtained, if possible, is typically not guaranteed to be globally optimal since there may be many local optima. It is hence of interest to extend the results of the paper to nonparametric instrumental quantile regression models, which is currently under investigation.

Acknowledgments

A major part of the work was done while the author was visiting the department of economics, MIT. He would like to thank Professor Victor Chernozhukov for his suggestions and encouragements, as well as Professor Yukitoshi Matsushita for his constructive comments.

References

- Ai, C., and Chen, X. (2003). Efficient estimation of conditional moment restrictions models containing unknown functions. *Econometrica* **71** 1795-1843.
- Belloni, A. and Chernozhukov, V. (2009a). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** 2011-2055.
- Belloni, A. and Chernozhukov, V. (2009b). Posterior inference in curved exponential families under increasing dimensions. [arXiv:0904.3132](#). Preprint.
- Bhatia, R. (1997). *Matrix Analysis*. Springer.
- Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises Theorem for discrete probability distributions. *Electron. J. Statist.* **3** 114-148.

- Blundell, R., Chen, X. and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613-1669.
- Bontemps, D. (2011). Bernstein-von Mises theorems for Gaussian regression with increasing number of regressions. *Ann. Statist.* **39** 2557-2584.
- Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems* **24** 1-19.
- Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903-923.
- Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica* **80** 277-321.
- Chen, X. and Reiss, M. (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* **27** 497-521.
- Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293-346.
- Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54-81.
- Clarke, B.S. and Ghosal, S. (2010). Reference priors for exponential families with increasing number dimension. *Electron. J. Statist.* **4** 737-780.
- Darolles, S., Fan, Y., Florens, J.P. and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541-1565.
- d'Haultfoeulle, X. (2011). On the completeness condition for nonparametric instrumental problems. *Econometric Theory* **27** 460-471.
- Florens, J.P. and Simoni, A. (2012a). Regularizing priors for linear inverse problems. *Scand. J. Statist.* **39** 458-475.
- Florens, J.P. and Simoni, A. (2012b). Nonparametric estimation of an instrumental variables regression: a quasi Bayesian approach based on a regularized posterior. *J. Econometrics* **170** 458-475.
- Gagliardini, P. and Scaillet, O. (2011). Nonparametric instrumental variables estimation of structural quantile effects. *Econometrica* **80** 1533-1562.
- Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer.
- Ghosal, S. (1999). Asymptotic normality of posterior distributions in high dimensional linear models. *Bernoulli* **5** 315-331.
- Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **73** 49-68.
- Ghosal, S., Ghosh, J. K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500-531.
- Ghosal, S. and van der Vaart, A.W. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192-223.
- Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^R metrics, $1 \leq R \leq \infty$. *Ann. Statist.* **39** 2883-2911.
- Hall, P. and Horowitz, J.L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904-2929.
- Hoffman, M. and Reiss, M. (2008). Nonlinear estimation for linear inverse problems with errors in operator. *Ann. Statist.* **36** 310-336.
- Horowitz, J.L. (2011). Applied nonparametric instrumental variables estimation.

- Econometrica* **79** 347-394.
- Horowitz, J.L. (2012). Specification testing in nonparametric instrumental variables estimation. *J. Econometrics* **167** 383-396.
- Horowitz, J.L. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* **75** 1191-1208.
- Härdle, W., Kerkycharian, F., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*. Springer.
- Jiang, W. and Tanner, M.A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207-2231.
- Johnstone, I.M. (2011). *Gaussian Estimation: Sequence and Multiresolution Models*. Unpublished draft.
- Kleijn, B.J.K. and van der Vaart, A.W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837-877.
- Knapik, B.T., van der Vaart, A.W. and van Zanten, J.H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626-2657.
- Kress, R. (1999). *Linear Integral Equations*. Second Edition. Springer.
- Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restriction models. *Ann. Statist.* **39** 3003-3033.
- Liu, J.S., Tian, L. and Wei, L.J. (2007). Implementation of estimating-function based inference procedures with Markov Chain Monte Carlo samplers. *J. Amer. Stat. Assoc.* **102** 881-888.
- Mallat, S. (2009). *A Wavelet Tour of Signal Processing*. Third Edition. Academic Press.
- Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863-884.
- Newey, W. and Powell, J. (2003). Instrumental variables estimation of nonparametric models. *Econometrica* **71** 1565-1578.
- Rudelson, M. (1999). Random vectors in the isotropic position. *J. Functional Anal.* **164** 60-72.
- Scriccio, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.* **34** 2897-2920.
- Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687-714.
- Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505-563.
- van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- Zhang, T. (2006a). From ϵ -entropy to KL entropy: analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180-2210.
- Zhang, T. (2006a). Information theoretical upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory* **54** 1307-1321.

Appendix A: Proofs of Lemmas 1, 3 and 4

Proof of Lemma 1. Part (ii) follows from Assumption 1 and the fact that $\{\phi_l\}$ is an orthonormal basis of $L_2[0, 1]$. Part (iii) follows from Rudelson's (1999) inequality and (i). For the reader's convenience, we state Rudelson's inequality in Appendix E. For Part (v), we first note that, by (iii) and Weyl's perturbation theorem (Bhatia, 1997, Problem III.6.13), $s_{\min}(\hat{\Phi}_{WX}) \geq \tau_{J_n} - O_P(\sqrt{J_n 2^{J_n}/n})$. Since now $\sqrt{J_n 2^{J_n}/n} = o(2^{-J_n r}) = o(\tau_{J_n})$, we have $s_{\min}(\hat{\Phi}_{WX}) \geq (1 - o_P(1))\tau_{J_n}$. For the proof of (i), denote by N the order of the Daubechies pair (φ, ψ) generating the CDV wavelet basis $\{\phi_l, l \geq 1\}$. Then, for each $x \in [0, 1]$ and each $j \geq J_0$, the number of nonzero elements in $\phi_{2^j+1}(x), \dots, \phi_{2^{j+1}}(x)$ is bounded by some constant depending only on N , and each $\phi_{2^j+k}(x)$ is bounded by some constant (depending only on ψ) times $2^{j/2}$ for all $k = 1, \dots, 2^j$. Similarly, $\phi_1, \dots, \phi_{2^{J_0}}$ are uniformly bounded. Therefore, there exists a constant D depending only on (φ, ψ) such that $\|\phi^J(x)\|_{\ell^2}^2 \leq D(2^{J_0} + \sum_{j=J_0}^{J-1} 2^j) = D2^J$ for all $x \in [0, 1]$.

Finally, we wish to show Part (iv). First, observe that $\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 \leq 2\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i] - \mathbb{E}[\phi^{J_n}(W)R]\|_{\ell^2}^2 + 2\|\mathbb{E}[\phi^{J_n}(W)R]\|_{\ell^2}^2$. By a simple moment calculation, the first term is $O_P(2^{J_n}/n)$. For the second term, by Assumptions 3 and 4 (ii),

$$\begin{aligned} \|\mathbb{E}[\phi^{J_n}(W)R]\|_{\ell^2}^2 &= \|\mathbb{E}[\phi^{J_n}(W)(g_0 - P_J g_0)(X)]\|_{\ell^2}^2 \\ &\lesssim \tau_{J_n}^2 \|g_0 - P_{J_n} g_0\|^2 \\ &\lesssim \tau_{J_n}^2 2^{-2J_n s}. \end{aligned}$$

This completes the proof. \square

Proof of Lemma 2. Step 1. We first show that $|\Sigma_n| = 1 + o(1)$ ($|\Sigma_n|$ denotes the determinant of Σ_n). Let $\lambda_{\min, n}$ and $\lambda_{\max, n}$ denote the minimum and maximum eigenvalues of Σ_n , respectively. Then, by Weyl's perturbation theorem, $1 - o(k_n^{-1}) \leq \lambda_{\min, n} \leq \lambda_{\max, n} \leq 1 + o(k_n^{-1})$, so that $(1 - o(k_n^{-1}))^{k_n} = \lambda_{\min, n}^{k_n} \leq |\Sigma_n| \leq \lambda_{\max, n}^{k_n} = (1 + o(k_n^{-1}))^{k_n}$. Here both sides converge to 1.

Step 2. By Step 1, we have

$$\begin{aligned} &\int |dN(0, \Sigma_n)(x) - dN(0, I_{k_n})(x)| dx \\ &= \frac{1}{(2\pi)^{k_n/2}} \int \left| \frac{1}{|\Sigma_n|^{1/2}} e^{-x^T \Sigma_n^{-1} x/2} - e^{-x^T x/2} \right| dx \\ &\leq \left| \frac{1}{|\Sigma_n|^{1/2}} - 1 \right| + \frac{1}{(2\pi)^{k_n/2} |\Sigma_n|^{1/2}} \int |e^{-x^T \Sigma_n^{-1} x/2} - e^{-x^T x/2}| dx \\ &\leq o(1) + \frac{1}{(2\pi)^{k_n/2} (1 + o(1))} \int e^{-x^T x/2} |e^{-x^T (\Sigma_n^{-1} - I_{k_n}) x/2} - 1| dx \end{aligned}$$

By assumption, we have $\epsilon_n := \|\Sigma_n^{-1} - I_{k_n}\|_{\text{op}} \leq \|\Sigma_n^{-1}\|_{\text{op}} \|I_{k_n} - \Sigma_n\|_{\text{op}} = o(k_n^{-1})$. Now, $|e^{-x^T (\Sigma_n^{-1} - I_{k_n}) x/2} - 1| \leq e^{\epsilon_n x^T x/2} - e^{-\epsilon_n x^T x/2}$. By a direct calculation, the conclusion follows from the fact that $(1 \pm \epsilon_n)^{k_n} = 1 + o(1)$. \square

Proof of Lemma 4. The first assertion follows from the assumption. Suppose now that \hat{A}_n is non-singular. Then, $\hat{A}_n^{-1}A_n = (\hat{A}_n - A_n + A_n)^{-1}A_n = (A_n^{-1}\hat{A}_n - I_{k_n} + I_{k_n})^{-1}$. Here, $A_n^{-1}\hat{A}_n - I_{k_n} = A_n^{-1}(\hat{A}_n - A_n)$, so that $\|A_n^{-1}\hat{A}_n - I_{k_n}\|_{\text{op}} \leq \|A_n^{-1}\|_{\text{op}}\|\hat{A}_n - A_n\|_{\text{op}} = s_{\min}^{-1}(A_n)\|\hat{A}_n - A_n\|_{\text{op}} = O_P(\epsilon_n^{-1}\delta_n)$. Let $\hat{\Delta} = I_{k_n} - A_n^{-1}\hat{A}_n$. Then, $\hat{A}_n^{-1}A_n = (I_{k_n} - \hat{\Delta})^{-1} = I_{k_n} + \sum_{m=1}^{\infty} \hat{\Delta}^m$ (Neumann series). Therefore, we conclude that $\|\hat{A}_n^{-1}A_n - I_{k_n}\|_{\text{op}} = \|\sum_{m=1}^{\infty} \hat{\Delta}^m\|_{\text{op}} \leq \sum_{m=1}^{\infty} \|\hat{\Delta}\|_{\text{op}}^m = \|\hat{\Delta}\|_{\text{op}} \cdot \sum_{m=0}^{\infty} \|\hat{\Delta}\|_{\text{op}}^m = O_P(\epsilon_n^{-1}\delta_n)$. \square

Appendix B: Proof of Theorem 2

We first prove the following lemma.

Lemma 5. *Suppose that the conditions of Theorem 2 are satisfied. Then, there exists a constant $D > 0$ such that*

$$\mathbb{P}\left\{\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2} > D\sqrt{2^{J_n}/n}\right\} \rightarrow 0.$$

Remark 8. It is standard to show that $\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2} = O_P(\sqrt{2^{J_n}/n})$, which, however, does not lead to the conclusion of Lemma 5 since the former only implies that for every sequence $M_n \rightarrow \infty$, $\mathbb{P}\{\|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2} > M_n\sqrt{2^{J_n}/n}\} \rightarrow 0$. Hence an additional step is needed. The current proof uses a truncation argument and Talagrand's concentration inequality.

Proof of Lemma 5. For a given $\lambda > 0$, define $U_i^- = U_i 1(|U_i| \leq \lambda)$ and $U_i^+ = U_i 1(|U_i| > \lambda)$. Since $0 = \mathbb{E}[U | W] = \mathbb{E}[U^- | W] + \mathbb{E}[U^+ | W]$, we have $\mathbb{E}_n[\phi^{J_n}(W_i)U_i] = n^{-1} \sum_{i=1}^n \{\phi^{J_n}(W_i)U_i^- - \mathbb{E}[\phi^{J_n}(W)U^-]\} + n^{-1} \sum_{i=1}^n \{\phi^{J_n}(W_i)U_i^+ - \mathbb{E}[\phi^{J_n}(W)U^+]\}$, by which we have

$$\begin{aligned} \|\mathbb{E}_n[\phi^{J_n}(W_i)U_i]\|_{\ell^2} &\leq \|n^{-1} \sum_{i=1}^n \{\phi^{J_n}(W_i)U_i^- - \mathbb{E}[\phi^{J_n}(W)U^-]\}\|_{\ell^2} \\ &\quad + \|n^{-1} \sum_{i=1}^n \{\phi^{J_n}(W_i)U_i^+ - \mathbb{E}[\phi^{J_n}(W)U^+]\}\|_{\ell^2} \\ &=: I + II. \end{aligned}$$

First, by Markov's inequality, we have for all $z > 0$,

$$\begin{aligned} \mathbb{P}(II > z) &\leq \frac{\mathbb{E}[II^2]}{z^2} \leq \frac{\sum_{l=1}^{2^{J_n}} \mathbb{E}[(\phi_l(W)U^+)^2]}{nz^2} \\ &\leq \frac{\sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) | W = w] \times \sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2]}{nz^2} \\ &\leq \frac{C_1 2^{J_n}}{nz^2} \times \sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) | W = w], \end{aligned}$$

where we have used that $\sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2] = \text{tr}(\Phi_{WW}) \leq 2^{J_n} s_{\max}(\Phi_{WW}) \leq C_1 2^{J_n}$ by Lemma 1 (ii). Thus, we have

$$\mathbb{P}\{II > \sqrt{C_1 2^{J_n}/n}\} \leq \sup_{w \in [0,1]} \mathbb{E}[U^2 1(|U| > \lambda) | W = w].$$

By assumption, the right side goes to zero as $\lambda \rightarrow \infty$.

Second, let $Z_i = \phi^{J_n}(W_i)U_i^- - \mathbb{E}[\phi^{J_n}(W)U^-]$ (denote by Z the generic version of Z_i). Let $\mathbb{S}^{2^{J_n}-1} := \{\alpha^{J_n} \in \mathbb{R}^{2^{J_n}} : \|\alpha^{J_n}\|_{\ell^2} = 1\}$. Then,

$$\begin{aligned} I &= \|\mathbb{E}_n[Z_i]\|_{\ell^2} \\ &= \sup_{\alpha^{J_n} \in \mathbb{S}^{2^{J_n}-1}} \mathbb{E}_n[(\alpha^{J_n})^T Z_i]. \end{aligned}$$

We make use of Talagrand's concentration inequality to bound the tail probability of I . For any $\alpha^{J_n} \in \mathbb{S}^{2^{J_n}-1}$, by Lemma 1, we have

$$\begin{aligned} \mathbb{E}[\{(\alpha^{J_n})^T Z\}^2] &\leq \sup_{w \in [0,1]} \mathbb{E}[U^2 \mid W = w] \times s_{\max}(\Phi_{WW}) \leq C_1^2, \\ |(\alpha^{J_n})^T Z| &\leq \lambda \sup_{w \in [0,1]} \|\phi^{J_n}(w)\|_{\ell^2} \leq D_1 \lambda \sqrt{2^{J_n}}, \text{ and} \\ (\mathbb{E}[I])^2 &\leq \mathbb{E}[I^2] \leq n^{-1} \sup_{w \in [0,1]} \mathbb{E}[U^2 \mid W = w] \times \sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2] \leq C_1^2 2^{J_n}/n, \end{aligned}$$

where $D_1 > 0$ is a constant. Thus, by Talagrand's inequality (see Theorem 5 in Appendix E), we have for all $z > 0$,

$$\mathbb{P}\{I \geq D_2(\sqrt{2^{J_n}/n} + \sqrt{z/n} + z\lambda\sqrt{2^{J_n}/n})\} \leq e^{-z},$$

where $D_2 > 0$ is a constant independent of λ and z .

The final conclusion follows from taking $\lambda = \lambda_n \rightarrow \infty$ and $z = z_n \rightarrow \infty$ sufficiently slowly. \square

Proof of Theorem 2. Let D_1, D_2 be some positive constants of which the values are understood in the context. For either $g_0 \in B_{\infty,\infty}^s$ or $B_{2,2}^s$, $\|g_0 - P_{J_n}g_0\| = O(2^{-J_n s}) = o(1)$, by which we have

$$\begin{aligned} &\sum_{l=1}^{2^{J_n}} \text{Var}\{\mathbb{E}_n[\phi_l(W_i)(g_0 - P_{J_n}g_0)(X_i)]\} \\ &\leq n^{-1} \sum_{l=1}^{2^{J_n}} \mathbb{E}[\phi_l(W)^2 \{(g_0 - P_{J_n}g_0)(X)\}^2] \\ &= n^{-1} \sum_{l=1}^{2^{J_n}} \iint \phi_l(w)^2 \{(g_0 - P_{J_n}g_0)(x)\}^2 f_{X,W}(x, w) dx dw \\ &\leq n^{-1} C_1 \|g_0 - P_{J_n}g_0\|^2 \times \sum_{l=1}^{2^{J_n}} \int \phi_l(w)^2 dw \\ &= o(2^{J_n}/n). \end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_n[\phi^{J_n}(W_i)R_i] &= \mathbb{E}_n[\phi^{J_n}(W_i)U_i] \\ &\quad + \mathbb{E}[\phi^{J_n}(W)(g_0 - P_n g_0)(X)] + \text{Rem},\end{aligned}$$

with $\|\text{Rem}\|_{\ell^2} = o_P(\sqrt{2^{J_n}/n})$. The second term on the right side is $O(\tau_{J_n} 2^{-J_n s})$ in the Euclidean norm. Together with Lemma 5, we have

$$\mathbb{P}\left\{\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 > D_1(\tau_{J_n}^2 2^{-2J_n s} + 2^{J_n}/n)\right\} \rightarrow 0.$$

Furthermore, by Lemma 1, we have

$$\text{tr}(\hat{\Phi}_{WW}) \leq 2^{J_n} s_{\max}(\hat{\Phi}_{WW}) \leq C_1(1 + o_P(1))2^{J_n}.$$

Taking these together, we have

$$\mathbb{P}\left[\|\mathbb{E}_n[\phi^{J_n}(W_i)R_i]\|_{\ell^2}^2 + n^{-1} \text{tr}(\hat{\Phi}_{WW}) \leq D_2(\tau_{J_n}^2 2^{-2J_n s} + 2^{J_n}/n)\right] \rightarrow 1.$$

By the proof of Theorem 1, this leads to the desired conclusion. \square

Appendix C: Proof of Theorem 3

For the notational convenience, define

$$\mathbb{E}_{\Pi_n}[\cdot | \mathcal{D}_n] := \int \cdot \Pi_n(dg | \mathcal{D}_n), \quad \mathbb{E}_{\tilde{\Pi}_n}[\cdot | \mathcal{D}_n] := \int \cdot \tilde{\Pi}_n(db^{J_n} | \mathcal{D}_n).$$

Proof of Theorem 3. Define the event

$$\mathcal{E}_{3n} = \{\mathcal{D}_n : \hat{\Phi}_{WX} \text{ and } \hat{\Phi}_{WW} \text{ are non-singular}\}.$$

Then, by Lemma 1, $\mathbb{P}\{1(\mathcal{E}_{3n}) = 1\} = \mathbb{P}(\mathcal{E}_{3n}) \rightarrow 1$. Suppose that $1(\mathcal{E}_{3n}) = 1$. Then, by (13), $\ell_{b^{J_n}}(\mathcal{D}_n)$ defined in the proof of Proposition 5 is bounded from below by $\hat{c}\|b^{J_n}\|_{\ell^2}^2$ + a term independent of b^{J_n} for some positive random variable \hat{c} . Hence, the integral $\mathbb{E}_{\tilde{\Pi}_n}[\|b^{J_n}\|_{\ell^2} | \mathcal{D}_n]$ is finite as soon as $1(\mathcal{E}_{3n}) = 1$. This proves the first assertion.

In what follows, we wish to prove the convergence rate result (11). First of all, by the triangle inequality and Jensen's inequality,

$$\begin{aligned}1(\mathcal{E}_{3n})\|\hat{g}_{QB} - g_0\| &\leq 1(\mathcal{E}_{3n})\|\hat{g}_{QB} - P_J g_0\| + \|g_0 - P_{J_n} g_0\| \\ &= 1(\mathcal{E}_{3n})\|\mathbb{E}_{\Pi_n}[g - P_J g_0 | \mathcal{D}_n]\| + \|g_0 - P_{J_n} g_0\| \\ &= 1(\mathcal{E}_{3n})\|\mathbb{E}_{\tilde{\Pi}_n}[b^{J_n} - b_0^{J_n} | \mathcal{D}_n]\|_{\ell^2} + \|g_0 - P_{J_n} g_0\| \\ &\leq 1(\mathcal{E}_{3n})\mathbb{E}_{\tilde{\Pi}_n}[\|b^{J_n} - b_0^{J_n}\|_{\ell^2} | \mathcal{D}_n] + \|g_0 - P_{J_n} g_0\|.\end{aligned}$$

Since $\|g_0 - P_{J_n} g_0\| = O(2^{-J_n s})$, it suffices to show that there exists a constant $M > 0$ such that

$$\begin{aligned}\mathbb{P}\left[1(\mathcal{E}_{3n})\mathbb{E}_{\tilde{\Pi}_n}[\|b^{J_n} - b_0^{J_n}\|_{\ell^2} | \mathcal{D}_n]\right. \\ \left.\leq M \max\{2^{-J_n s}, 2^{J_n r} \sqrt{2^{J_n}/n}, 2^{J_n r} \epsilon_n \varrho_n (\log n)^{1/2}\}\right] \rightarrow 1.\end{aligned}$$

Let $\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)$ be the (random) density defined in the proof of Theorem 1. Note that $\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)$ is well-defined as soon as $1(\mathcal{E}_{3n}) = 1$. Let $\delta_n := \epsilon_n + \tau_{J_n} 2^{-J_n s}$. Then we have:

Lemma 6. *There exists a constant $c_1 > 0$ such that for every sequence $M_n \rightarrow \infty$ with $M_n = o(L_n)$,*

$$\mathbb{P} \left\{ 1(\mathcal{E}_{3n}) \int \|\theta^{J_n}\|_{\ell^2} \cdot |\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})| d\theta^{J_n} \leq e^{-c_1 M_n n \delta_n^2} + M_n \sqrt{n} \delta_n \varrho_n \right\} \rightarrow 1,$$

where $\tilde{\Delta}_n := \sqrt{n} \mathbb{E}_n[\phi^{J_n}(W_i) R_i]$.

We defer the proof of Lemma 6 after the proof of this theorem. Here we have

$$\begin{aligned} & 1(\mathcal{E}_{3n}) \left[\int \|\theta^{J_n}\|_{\ell^2} dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n}) d\theta^{J_n} \right]^2 \\ & \leq 1(\mathcal{E}_{3n}) \int \|\theta^{J_n}\|_{\ell^2}^2 dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n}) d\theta^{J_n} \\ & \leq \|\tilde{\Delta}_n\|_{\ell^2}^2 + \text{tr}(\hat{\Phi}_{WW}). \end{aligned}$$

By the proof of Theorem 2, there exists a constant $D_1 > 0$ such that $\mathbb{P}\{\|\tilde{\Delta}_n\|_{\ell^2}^2 + \text{tr}(\hat{\Phi}_{WW}) \leq D_1(n\tau_{J_n}^2 2^{-2J_n s} + 2^{J_n})\} \rightarrow 1$. Thus, for every sequence $M_n \rightarrow \infty$ with $M_n = o(L_n)$, with probability approaching one,

$$\begin{aligned} & \sqrt{D_1(n\tau_{J_n}^2 2^{-2J_n s} + 2^{J_n})} + e^{-c_1 M_n n \delta_n^2} + M_n \sqrt{n} \delta_n \varrho_n \\ & \geq 1(\mathcal{E}_{3n}) \int \|\theta^{J_n}\|_{\ell^2} \pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) \\ & = 1(\mathcal{E}_{3n}) \sqrt{n} \int \|\hat{\Phi}_{WX}(b^{J_n} - b_0^{J_n})\|_{\ell^2} \tilde{\pi}_n(b^{J_n} \mid \mathcal{D}_n) db^{J_n} \\ & \geq 1(\mathcal{E}_{3n}) \sqrt{n} s_{\min}(\hat{\Phi}_{WX}) \mathbb{E}_{\tilde{\Pi}_n}[\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \mid \mathcal{D}_n]. \end{aligned}$$

Take $M_n \rightarrow \infty$ sufficiently slowly such that $M_n = O((\log n)^{1/2})$. Since the left side is then $\lesssim \max\{\sqrt{n} \tau_{J_n} 2^{-J_n s}, \sqrt{2^{J_n}}, \sqrt{n} \epsilon_n \varrho_n (\log n)^{1/2}\}$, there exists a constant $D_2 > 0$ such that

$$\begin{aligned} & \mathbb{P} \left[1(\mathcal{E}_{3n}) s_{\min}(\hat{\Phi}_{WX}) \mathbb{E}_{\tilde{\Pi}_n}[\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \mid \mathcal{D}_n] \leq D_2 \max\{\tau_{J_n} 2^{-J_n s}, \sqrt{2^{J_n}/n}, \epsilon_n \varrho_n (\log n)^{1/2}\} \right] \rightarrow 1. \end{aligned}$$

Finally, by Lemma 1, $\mathbb{P}(s_{\min}(\hat{\Phi}_{WX}) \geq 0.5 \tau_{J_n}) \rightarrow 1$, by which we have

$$\begin{aligned} & \mathbb{P} \left[1(\mathcal{E}_{3n}) \mathbb{E}_{\tilde{\Pi}_n}[\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \mid \mathcal{D}_n] \leq 2D_2 \max\{2^{-J_n s}, \tau_{J_n}^{-1} \sqrt{2^{J_n}/n}, \tau_{J_n}^{-1} \epsilon_n \varrho_n (\log n)^{1/2}\} \right] \rightarrow 1. \end{aligned}$$

This leads to the desired conclusion. \square

Proof of Lemma 6. As before, we say that a sequence of random variables A_n is *eventually* bounded by another sequence of random variables B_n if $\mathbb{P}(A_n \leq B_n) \rightarrow 1$.

Take any $M_n \rightarrow \infty$ with $M_n = o(L_n)$. Then,

$$\begin{aligned}
& 1(\mathcal{E}_{3n}) \int \|\theta^{J_n}\|_{\ell^2} \cdot |\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})| d\theta^{J_n} \\
& \leq 1(\mathcal{E}_{3n}) \int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} \|\theta^{J_n}\|_{\ell^2} \cdot |\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) - dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})| d\theta^{J_n} \\
& \quad + 1(\mathcal{E}_{3n}) \int_{\|\theta^{J_n}\|_{\ell^2} > M_n \sqrt{n} \delta_n} \|\theta^{J_n}\|_{\ell^2} \pi_n^*(\theta^{J_n} \mid \mathcal{D}_n) d\theta^{J_n} \\
& \quad + 1(\mathcal{E}_{3n}) \int_{\|\theta^{J_n}\|_{\ell^2} > M_n \sqrt{n} \delta_n} \|\theta^{J_n}\|_{\ell^2} dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n}) d\theta^{J_n} \\
& =: I + II + III.
\end{aligned}$$

We divide the rest of the proof into three steps.

Step 1. Claim: there exists a constant $c_2 > 0$ such that $\mathbb{P}(II \leq e^{-c_2 M_n^2 n \delta_n^2}) \rightarrow 1$.

(Proof of Step 1): The assertion of Step 1 follows from the same line as in the proof of Proposition 5 by noting that for any $c > 0$, $x e^{-cx^2} \leq e^{-cx^2/2}$ for all $x > 0$ sufficiently large. Hence the proof is omitted.

Step 2. Claim: there exists a constant $c_3 > 0$ such that $\mathbb{P}(III \leq e^{-c_3 M_n n \delta_n^2}) \rightarrow 1$.

(Proof of Step 2): By the Cauchy-Schwarz inequality, the square of III is bounded by $\int \|\theta^{J_n}\|_{\ell^2}^2 dN(\tilde{\Delta}_n, \hat{\Phi}_{WW}) d\theta^{J_n} \int_{\|\theta^{J_n}\|_{\ell^2} > M_n \sqrt{n} \delta_n} dN(\tilde{\Delta}_n, \hat{\Phi}_{WW}) d\theta^{J_n}$. Here the first integral is bounded by

$$\|\tilde{\Delta}_n\|_{\ell^2}^2 + \text{tr}(\hat{\Phi}_{WW}),$$

which is eventually bounded by $D(n\tau_{J_n}^2 2^{-J_n s} + 2^{J_n})$ for some constant $D > 0$ by the proof of Theorem 2. On the other hand, by the proof of Theorem 1, the second integral is eventually bounded by $\int_{\|\theta^{J_n}\|_{\ell^2} > \sqrt{M_n n} \delta_n} dN(0, I_{2^{J_n}}) d\theta^{J_n}$. By Borell's inequality for Gaussian measures (see (20)), the last integral is bounded by $e^{-c' M_n n \delta_n^2}$ for some small constant $c' > 0$. Taking these together, we obtain the conclusion of Step 2 by choosing the constant $c_3 > 0$ sufficiently small.

Step 3. Claim: there exists a constant $c_4 > 0$ such that $\mathbb{P}(I \leq e^{-c_4 M_n^2 n \delta_n^2} + M_n \sqrt{n} \delta_n \varrho_n) \rightarrow 1$.

(Proof of Step 3): Let $\mathcal{C}_n := \{\theta^{J_n} \in \mathbb{R}^{2^{J_n}} : \|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n\}$. Let $\pi_{n, \mathcal{C}_n}^*(\theta^{J_n} \mid \mathcal{D}_n)$ and $dN^{\mathcal{C}_n}(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ denote the probability densities obtained by first restricting $\pi_n^*(\theta^{J_n} \mid \mathcal{D}_n)$ and $dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ to the ball \mathcal{C}_n and then renormalizing, respectively. Then, abbreviating $\pi_n^*(d\theta^{J_n} \mid \mathcal{D}_n)$ by π_n^* , $\pi_{n, \mathcal{C}_n}^*(d\theta^{J_n} \mid \mathcal{D}_n)$ by π_{n, \mathcal{C}_n}^* , $dN(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$ by dN , and $dN^{\mathcal{C}_n}(\tilde{\Delta}_n, \hat{\Phi}_{WW})(\theta^{J_n})$

by dN^{C_n} we have

$$\begin{aligned}
I &\leq 1(\mathcal{E}_{3n}) \int \|\theta^{J_n}\|_{\ell^2} \cdot |\pi_{n,C_n}^* - dN^{C_n}| \\
&\quad + 1(\mathcal{E}_{3n}) \int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} \|\theta^{J_n}\|_{\ell^2} \cdot |\pi_{n,C_n}^* - \pi_n^*| \\
&\quad + 1(\mathcal{E}_{3n}) \int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} \|\theta^{J_n}\|_{\ell^2} \cdot |dN^{C_n} - dN| \\
&=: IV + V + IV.
\end{aligned}$$

By the proof of Theorem 1, the term IV is eventually bounded by

$$1(\mathcal{E}_{3n}) M_n \sqrt{n} \delta_n \int |\pi_{n,C_n}^* - dN^{C_n}| \leq M_n \sqrt{n} \delta_n \varrho_n.$$

For the term V , we have

$$\begin{aligned}
V &\leq 1(\mathcal{E}_{3n}) M_n \sqrt{n} \delta_n \int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} |\pi_{n,C_n}^* - \pi_n^*| \\
&= 1(\mathcal{E}_{3n}) M_n \sqrt{n} \delta_n \times \frac{\int_{\|\theta^{J_n}\|_{\ell^2} > M_n \sqrt{n} \delta_n} \pi_n^*}{\int_{\|\theta^{J_n}\|_{\ell^2} \leq M_n \sqrt{n} \delta_n} \pi_n^*}.
\end{aligned}$$

By the proof of Proposition 5, there exists a constant $c_5 > 0$ such that the ratio of the integrals on the right side is eventually bounded by $e^{-c_5 M_n^2 n \delta_n^2}$, so that $\mathbb{P}(V \leq e^{-c_5 M_n^2 n \delta_n^2 / 2}) \rightarrow 1$. Likewise, by Borell's inequality for Gaussian measures, there exists a constant $c_6 > 0$ such that $\mathbb{P}(VI \leq e^{-c_6 M_n n \delta_n^2}) \rightarrow 1$. Taking these together, we obtain the conclusion of Step 3 by choosing the constant $c_4 > 0$ sufficiently small.

Finally, Steps 1-3 lead to the conclusion of Lemma 6. \square

Appendix D: Proofs for Section 5

Proof of Proposition 2. For either case of product or isotropic priors, it suffices to check conditions P1) and P2) in Theorem 1. We shall do this with the choice $\epsilon_n = \sqrt{2^{J_n} (\log n) / n} \sim (\log n)^{1/2} n^{-(r+s)/(2r+2s+1)}$.

Case of product priors: Let $c_{\min} := \min_{x \in [-A, A]} q(x) > 0$. Since $\|b^{J_n} - b_0^{J_n}\|_{\ell^2}^2 = \sum_{l=1}^{2^{J_n}} (b_l - b_{0l})^2 \leq 2^{J_n} \max_{1 \leq l \leq 2^{J_n}} (b_l - b_{0l})^2$, we have

$$\begin{aligned}
\tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq \epsilon_n) &\geq \tilde{\Pi}_n \left(b^{J_n} : \max_{1 \leq l \leq 2^{J_n}} |b_l - b_{0l}| \leq \epsilon_n / \sqrt{2^{J_n}} \right) \\
&\geq \prod_{l=1}^{2^{J_n}} \tilde{\Pi}_n(b^{J_n} : |b_l - b_{0l}| \leq \epsilon_n / \sqrt{2^{J_n}}).
\end{aligned}$$

Since $\exists \epsilon \in (0, A)$, $b_{0l} \in [-A + \epsilon, A - \epsilon]$ for all $l \geq 1$, for all n sufficiently large, the last expression is bounded from below by

$$\left(\frac{c_{\min} \epsilon_n}{\sqrt{2^{J_n}}} \right)^{2^{J_n}} = e^{-2^{J_n} \log(\sqrt{2^{J_n}} / (c_{\min} \epsilon_n))} \geq e^{-C n \epsilon_n^2},$$

where $C > 0$ is a sufficiently large constant, which verifies condition P1).

Second, with this ϵ_n , γ_n in condition P2) is $\sim (\log n)^{1/2} n^{-s/(2r+2s+1)}$. Let, say, $L_n \sim (\log n)^{1/2}$ so that $L_n \gamma_n \sim (\log n) n^{-s/(2r+2s+1)}$. Then, $\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n\} \subset [-A, A]^{2^{J_n}}$ for all n sufficiently large, so that $\tilde{\pi}_n(b^{J_n}) = \prod_{l=1}^{2^{J_n}} q(b_l)$ is positive for all $\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n$. Let $\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n$ and $\|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n$. Then,

$$\begin{aligned} \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} &= \exp \left[\sum_{l=1}^{2^{J_n}} \{ \log q(b_{0l} + b_l) - \log q(b_{0l} + \tilde{b}_l) \} \right] \\ &\leq \exp \left\{ L \sum_{l=1}^{2^{J_n}} |b_l - \tilde{b}_l| \right\} \\ &\leq \exp \left\{ L \sqrt{2^{J_n}} \|b^{J_n} - \tilde{b}^{J_n}\|_{\ell^2} \right\} \\ &\leq e^{2L \sqrt{2^{J_n}} L_n \gamma_n} = e^{o(1)}, \end{aligned}$$

where the last step is due to $s > 1/2$. Likewise, we have

$$\frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} \geq e^{-2L \sqrt{2^{J_n}} L_n \gamma_n} = e^{-o(1)}.$$

Therefore, condition P2) is verified.

Case of isotropic priors: Let $c_{\min} := \min_{x \in [0, A]} r(x) > 0$. Then, for all n sufficiently large,

$$\begin{aligned} \tilde{\Pi}_n(b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq \epsilon_n) &= \frac{\int_{\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq \epsilon_n} r(\|b^{J_n}\|_{\ell^2}) db^{J_n}}{\int r(\|b^{J_n}\|_{\ell^2}) db^{J_n}} \\ &= \frac{\int_{\|b^{J_n}\|_{\ell^2} \leq \epsilon_n} r(\|b^{J_n} + b_0^{J_n}\|_{\ell^2}) db^{J_n}}{\int r(\|b^{J_n}\|_{\ell^2}) db^{J_n}} \\ &\geq \frac{c_{\min} \int_{\|b^{J_n}\|_{\ell^2} \leq \epsilon_n} db^{J_n}}{\int r(\|b^{J_n}\|_{\ell^2}) db^{J_n}} \\ &= c_{\min} \frac{\int_{x \in [0, \epsilon_n]} x^{2^{J_n}-1} dx}{\int_0^\infty x^{2^{J_n}-1} r(x) dx} \\ &\geq c_{\min} \left(\frac{\epsilon_n}{2^{J_n}} \right)^{2^{J_n}} \times e^{-c 2^{J_n} \log(2^{J_n})} \\ &= c_{\min} e^{-2^{J_n} \log(2^{J_n}/\epsilon_n) - c 2^{J_n} \log(2^{J_n})} \\ &\geq e^{-C n \epsilon_n^2}, \end{aligned}$$

where $C > 0$ is a sufficiently large constant, which verifies condition P1).

Second, with this ϵ_n, γ_n in condition P2) is $\sim (\log n)^{1/2} n^{-s/(2r+2s+1)}$. Let $L_n \sim (\log n)^{1/2}$ so that $L_n \gamma_n \sim (\log n) n^{-s/(2r+2s+1)}$. Since $\|b_0^{J_n}\|_{\ell^2} \leq \|g_0\| < A$ and $L_n \gamma_n \rightarrow 0$, $\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n\} \subset \{b^{J_n} : \|b^{J_n}\|_{\ell^2} \leq A\}$ for all n sufficiently large, so that $\tilde{\pi}_n(b^{J_n}) \propto r(\|b^{J_n}\|_{\ell^2})$ is positive for all $\|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n$. Let $\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n$ and $\|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n$. Then, by Bessel's inequality,

$$\|b_0^{J_n} + b^{J_n}\|_{\ell^2} \leq \|b_0^{J_n}\|_{\ell^2} + L_n \gamma_n \rightarrow \|g_0\|,$$

and likewise we have

$$\|b_0^{J_n} + b^{J_n}\|_{\ell^2} \geq \|b_0^{J_n}\|_{\ell^2} - L_n \gamma_n \rightarrow \|g_0\|.$$

Therefore, we conclude that

$$\begin{aligned} \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} &= \frac{r(\|b_0^{J_n} + b^{J_n}\|_{\ell^2})}{r(\|b_0^{J_n} + \tilde{b}^{J_n}\|_{\ell^2})} \\ &\rightarrow \frac{r(\|g_0\|)}{r(\|g_0\|)} = 1, \end{aligned}$$

uniformly in $\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n$ and $\|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n$, which verifies condition P2). \square

Proof of Proposition 3. Given the proof of Proposition 2 and the discussion following Theorem 3, it is sufficient to verify that ϱ_n is $O((\log n)^{-1})$. However, this is readily verified by tracking the proof of Proposition 2. \square

Proof of Proposition 4. The proof is similar in spirit to that of Proposition 2. Hence we only give a sketch of the proof.

Case (a): Condition P1) is verified with $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$. Then, γ_n in P2) is $\sim (\log n)^{1/2} n^{-s/(2r+2s+1)} \sim (\log n)^{1/2} 2^{-J_n s}$. Because $\tilde{\pi}_n$ is constant on the support, condition P2) is verified if the support of $\tilde{\pi}_n$ contains the ball $\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n\}$ for all n sufficiently large for some $L_n \rightarrow \infty$. Let, say, $L_n \sim (\log n)^{1/4}$, so that $L_n \gamma_n \sim (\log n)^{3/4} 2^{-J_n s}$. Since $\{b^{J_n} : \|b^{J_n} - b_0^{J_n}\|_{\ell^2} \leq L_n \gamma_n\} \subset \{b^{J_n} : \max_{1 \leq l \leq 2^{J_n}} |b_l - b_{0l}| \leq L_n \gamma_n\}$, condition P2) is verified if $L_n \gamma_n = o(A_n 2^{-J_n(s+1/2)})$. This is satisfied since $A_n 2^{-J_n(s+1/2)} \sim (\log n) 2^{-J_n s}$. The second assertion follows because in this case $\varrho_n = 0$ for all n sufficiently large.

Case (b): Condition P1) is verified with $\epsilon_n \sim \sqrt{2^{J_n}(\log n)/n}$ without a significant difficulty. Then, γ_n in P2) is $\sim (\log n)^{1/2} n^{-s/(2r+2s+1)} \sim (\log n)^{1/2} 2^{-J_n s}$. Let, say, $L_n \sim (\log n)^{1/2}$. To establish the desired conclusion in this case, it is sufficient to prove that

$$\left| \log \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} \right| = O((\log n)^{-1}),$$

uniformly in $\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n$ and $\|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n$. Let $\|b^{J_n}\|_{\ell^2} \leq L_n \gamma_n$ and $\|\tilde{b}^{J_n}\|_{\ell^2} \leq L_n \gamma_n$. Define $a_1, \dots, a_{2^{J_n}}$ by $a_k = 1$ for $k = 1, \dots, 2^{J_0}$ and $a_{2^j+k} =$

$2^{j(s+1/2)}$ for $k = 1, \dots, 2^j; j = J_0, \dots, J_n - 1$. Then, by construction,

$$\begin{aligned} \log \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} &= \frac{1}{2A_n^2} \sum_{l=1}^{2^{J_n}} a_l^2 \{-(b_{0l} + b_l)^2 + (b_{0l} + \tilde{b}_l)^2\} \\ &= \frac{1}{2A_n^2} \sum_{l=1}^{2^{J_n}} a_l^2 \{-b_l^2 + \tilde{b}_l^2 - 2b_{0l}(b_l - \tilde{b}_l)\}. \end{aligned}$$

Observe that

$$\begin{aligned} \frac{1}{A_n^2} \sum_{l=1}^{2^{J_n}} a_l^2 b_l^2 &\leq \frac{a_{2^{J_n}}^2}{A_n^2} \sum_{l=1}^{2^{J_n}} b_l^2 \leq \frac{2^{J_n(2s+1)} L_n^2 \gamma_n^2}{A_n^2} \sim (\log n)^{-1}, \\ \left(\sum_{l=1}^{2^{J_n}} a_l^2 b_{0l} b_l \right)^2 &\leq \left(\sum_{l=1}^{2^{J_n}} a_l^2 b_{0l}^2 \right) \left(\sum_{l=1}^{2^{J_n}} a_l^2 b_l^2 \right) \leq D 2^{J_n} \left(\sum_{l=1}^{2^{J_n}} a_l^2 b_l^2 \right) \leq D 2^{J_n(2s+2)} L_n^2 \gamma_n^2, \end{aligned}$$

where $D > 0$ is constant depending only on $\|g_0\|_{2,s,s}$. The second inequality leads to that

$$\frac{1}{A_n^2} \left| \sum_{l=1}^{2^{J_n}} a_l^2 b_{0l} b_l \right| \leq \frac{\sqrt{D} 2^{J_n(s+1)} L_n \gamma_n}{A_n^2} \sim (\log n)^{-1}.$$

Therefore, we conclude that

$$\sup_{\|b^{J_n}\|_{\ell_2} \leq L_n \gamma_n, \|\tilde{b}^{J_n}\|_{\ell_2} \leq L_n \gamma_n} \left| \log \frac{\tilde{\pi}_n(b_0^{J_n} + b^{J_n})}{\tilde{\pi}_n(b_0^{J_n} + \tilde{b}^{J_n})} \right| = O((\log n)^{-1}).$$

This completes the proof. \square

Appendix E: Technical tools

We state here Rudelson's inequality for the reader's convenience.

Theorem 4 (Rudelson's (1999) inequality). *Let Z_1, \dots, Z_n be i.i.d. random vectors in \mathbb{R}^k with $\Sigma := E[Z_1^{\otimes 2}]$. Then, for all $k \geq e^2$,*

$$\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n Z_i^{\otimes 2} - \Sigma \right\|_{\text{op}} \right] \leq \max\{\|\Sigma\|_{\text{op}}^{1/2} \delta, \delta^2\}, \quad \delta = D \sqrt{\frac{\log k}{n} \mathbb{E}[\max_{1 \leq i \leq n} \|Z_i\|_{\ell_2}^2]},$$

where D is a universal constant.

Rudelson's inequality implies the following corollary useful in our application.

Corollary 1. Let $(X_1, Y_1^T)^T, \dots, (X_n, Y_n^T)^T$ be i.i.d. random vectors with $X_i \in \mathbb{R}^{k_1}, Y_i \in \mathbb{R}^{k_2}$, and $k_1 + k_2 \geq e^2$. Let $\Sigma_X := \mathbb{E}[X_1^{\otimes 2}]$, $\Sigma_Y := \mathbb{E}[Y_1^{\otimes 2}]$ and $\Sigma_{XY} := \mathbb{E}[X_1 Y_1^T]$. Suppose that there exists a finite number m such that $\mathbb{E}[\max_{1 \leq i \leq n} \|X_i\|_{\ell^2}^2] \vee \mathbb{E}[\max_{1 \leq i \leq n} \|Y_i\|_{\ell^2}^2] \leq m$. Then,

$$\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n X_i Y_i^T - \Sigma_{XY} \right\|_{\text{op}} \right] \leq \max\{(\|\Sigma_X\|_{\text{op}}^{1/2} \vee \|\Sigma_Y\|_{\text{op}}^{1/2})\delta, \delta^2\},$$

$$\text{with } \delta = D \sqrt{\frac{m \log(k_1 \vee k_2)}{n}},$$

where D is a universal constant.

Proof. Let $Z_i = (X_i, Y_i^T)^T$, and apply Rudelson's inequality to Z_1, \dots, Z_n . Note that by the variational characterization of the operator norm, we have $\|n^{-1} \sum_{i=1}^n X_i Y_i^T - \Sigma_{XY}\|_{\text{op}} \leq \|n^{-1} \sum_{i=1}^n Z_i^{\otimes 2} - \mathbb{E}[Z_1^{\otimes 2}]\|_{\text{op}}$, and by the Cauchy-Schwarz inequality, $\|\mathbb{E}[Z_1^{\otimes 2}]\|_{\text{op}} \leq 2\|\Sigma_X\|_{\text{op}} + 2\|\Sigma_Y\|_{\text{op}}$. \square

Finally, we recall celebrated Talagrand's (1996) concentration inequality for general empirical processes. The following version is due to Massart (2000). Here, for a generic class \mathcal{F} of measurable functions on some measurable space \mathcal{X} , we say that \mathcal{F} is pointwise measurable if there exists a countable class \mathcal{G} of measurable functions on \mathcal{X} such that for any $f \in \mathcal{F}$, there exists a sequence $\{g_m\} \subset \mathcal{G}$ with $g_m(x) \rightarrow f(x)$ for all $x \in \mathcal{X}$. See Chapter 2.3 of van der Vaart and Wellner (1996).

Theorem 5 (Massart's form of Talagrand's inequality). Let $\xi_i, i = 1, 2, \dots, n$ be i.i.d. random variables taking values in some measurable space \mathcal{X} . Let \mathcal{F} be a pointwise measurable class of functions on \mathcal{X} such that $\mathbb{E}[f(\xi_1)] = 0$ for all $f \in \mathcal{F}$ and $\sup_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} |f(x)| \leq B$ for some constant $B > 0$. Let σ^2 be any positive constant such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(\xi_1)]$. Let $Z := \sup_{f \in \mathcal{F}} |\sum_{i=1}^n f(\xi_i)|$. Then, for all $x > 0$, we have

$$\mathbb{P}\{Z \geq C(\mathbb{E}[Z] + \sigma\sqrt{nx} + Bx)\} \leq e^{-x},$$

where $C > 0$ is a universal constant.